HOW ARE JOB APPLICANTS DISADVANTAGED BY GENDER-BASED DOUBLE STANDARDS IN A NATURAL SETTING

A Dissertation
Presented to the Faculty of the Graduate School
of Cornell University
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy

by
Esther Quintero
August 2008

HOW ARE JOB APPLICANTS DISADVANTAGED BY GENDER BASED DOUBLE STANDARDS IN A NATURAL SETTING

Esther Quintero, Ph.D.
Cornell University 2008

Controlled experiments have found that mixed-sex interaction triggers the use of gender status beliefs encouraging actors to view men as more status worthy than women (Correll & Ridgeway 2003; Foschi 2000). When these mechanisms are at play in actual employment settings, the implication is that employers will be more inclined to hire, promote, and praise male workers even when they produce identical evidence of competence that their female counterparts.

Although hiring settings are rarely accessible to researchers, this project identified a unique exception that permitted direct observation and data collection on real evaluations made in the course of live interactions. The context of this study is the Spanish exam system that is used to recruit applicants to fill important and highly desirable government jobs where women are currently under-represented.

The first part of this project analyzes how male and female applicants fare at each testing round. I examined quantitative pass/fail exam data gathered from the Internet and information obtained from direct observation of exam sessions. Consistent with gender status theories, I found that female applicants scored higher than male applicants when exams evaluated female-typed skills and lower than male applicants when exams involved assessing neutral abilities – this effect was

substantial in less structured exams (i.e. exam 3) where judges and applicants engage in an actual dialogue.

The second section of this work analyzes the features and content of the live conversations that take place in the Q&A portion of exam 3. I audio-taped and transcribed 83 judge-applicant conversations and found that exam judges behave differently toward male and female applicants even when applicants perform at the same level. In particular, judges interrupt women more, ask them more questions, and are less persuaded by the objective quality of their answers.

The main contributions of this work are (i) a substantive one, of increasing our knowledge on the sources of gender segregation, (ii) applying theory developed in laboratory settings in a natural context, and (iii) a methodological one, of using the method of conversation analysis to understand the dynamics of judges-applicants interactions.

BIOGRAPHICAL SKETCH

Esther Quintero received a B.A. degree in History from the Universidad de Sevilla in 2001, and a M.A. degree in Sociology from Cornell University in 2006.

Esther has several graduate awards from the University of Seville and Cornell University. In 2001, she received a Graduate Fellowship for Study in the United States. In 2005, Esther was awarded the Luigi Einaudi Dissertation Fellowship and an International Travel Grant from the Institute of European Studies at Cornell.

The National Science Foundation awarded a Dissertation Improvement Grant to Professor Shelley Correll (Principal Investigator) and Esther Quintero (co-Principal Investigator) in January 2007.

In May 2008, Esther received the Robin M. Williams, Jr. Best Graduate Student Paper Award for "Gender-Based Double Standards for the Evaluation of Job Applicants in a Natural Setting."

Esther served as a Teaching Assistant at Cornell University between 2002 and 2004, and collaborated as a Research Assistant with Professor Gita Sen at the Indian Institute of Management in Bangalore, India in 2005.

In April 2008, Esther was offered a research position at a Madrid based R&D company specialized in intelligence innovation.

ACKNOWLEDGMENTS

This work was made possible with support from Cornell University's Institute of European Studies (Luigi Einaudi Fellowship 2005-2006; International Travel Grant 2005) and an NSF Dissertation Improvement Grant awarded to Shelley Correll (Principal Investigator) and Esther Quintero (co-PI) in 2007.

I wish to thank all members of my dissertation committee for their support throughout the last four years. I am particularly indebted to Shelley Correll, the chairperson of my special committee, and David Grusky, who was a member of my dissertation committee until May 2008. I am deeply grateful to Kim Weeden who joined the committee more recently but whose input helped me improve this work in substantive ways. Finally, I want to thank Celia Valiente and Mabel Berezin for their useful suggestions and steady encouragement.

My friends/colleagues Stephen Benard and Manuel Bagüés have been immensely helpful, I hope that I continue to learn from them in the years to come as they are academics that I particularly look up to. I would also like to thank the staff of Cornell University's Sociology Department, particularly Sharon Sandlan who has willingly assisted me with multiple requests necessary for the completion of this work.

This effort took several years; during this time many people have been important at levels that transcend pure academic work. I am particularly thankful to my friend, partner, and colleague César Talón who, in the last three years has been there for me through the good, the

bad, and the ugly, relentlessly showing unparalleled intellectual and personal support.

Some of my friends in Ithaca have been a great source of inspiration, especially Ed, Arnout, Matt, Cristina, Nathalie, and Manas among others. Back in Spain, my friends José Mª, Gloria, Jose, Tito, and Sonsoles have also been important to me.

Family always matters and this is no exception. I want to thank my parents, Isabel and Manolo, for the opportunities they made available for me; and Carlos, my brother, for loving me unconditionally. I am grateful to have shared so many fun moments with my aunts Luisa and Manola, for these have definitely helped me keep an even keel. The Talón-Bueno family (Mariví, Jesús, and Jesús Jr.) has also been a tremendous source of stability and strength since I arrived in Madrid.

The average person may not know what dissertations are or what they require. I used to think only those who "knew" were in a position to understand my work, thus understand me. In the last year I have come to embrace what each person was in a position to give me, for I am this work but far more than that. I am now fully aware that all forms of support and inspiration play an important role. My feeling, looking back, is one of immense gratitude to all these people…

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER ONE
INTRODUCTION

Gender inequality in employment persists despite important structural changes such as the movement of women into paid labor and the increase in the number of women pursuing higher education. Despite these steps forward, women are still disadvantaged in the job market. Wage differences in particular exist largely because men and women occupy different jobs (Reskin 1993). Although the causes of occupational sex segregation are multifaceted, discrimination by employers surely contributes to it (Blau & Kahn 2006).

When individuals deduce others' ability from their performances, it is often the case that not all actors are assessed according to the same criteria. Despite providing the same evidence of skill, actors are evaluated differently because ability inferences are based on performers' personal attributes such as race or gender (Foschi 1989). Frequently, individuals are judged based on generalized beliefs about the social category to which they belong, rather than on the objective quality of what they say/do. This process has important implications in a wide variety of settings, particularly those where implicit and explicit evaluations are the focus of the interaction – e.g. job interviews, promotions, salary decisions etc.

This work draws from status characteristics theory (SCT) and double standards theory (DST), both of which are concerned with how status attributes such as gender influence behavior and the evaluation of task performance in mixed-sex social interaction. SCT and DST are not theories of discrimination per se but, insofar as they explain how and why

employers come to prefer men when making hiring, promotion and salary decisions, these theories are relevant for understanding the persistence of gender inequalities in the labor market.

Even though sociologists and social psychologists have amply demonstrated that gender systematically organizes the perception of competence and influence in favor of men (see Correll & Ridgeway 2003; Foschi 2000), researchers have argued that the dynamics of gender discrimination in the workplace will not be fully understood until detailed data are collected on less scripted settings (Biernat & Fuegen 2001; Ridgeway & Correll 2004).

Real-world hiring contexts are seldom accessible to researchers, thus observation and/or systematic data collection are rare. In fact, Fernández and Sosa (2005) have noted how almost all research on gender segregation begins with data on people who already have jobs, and very little empirical evidence exists on the workings of hiring[1]. While Fernández and Sosa (2005) have data on real evaluations of employees, their research does not examine the processes by which evaluations unfold. My work takes advantage of an original context where I was able to observe and collect data on the actual interactions taking place as potential employees were interviewed by exam evaluators.

Context & Setting

The context of this study is the exam system used in Spain to recruit applicants for important and highly desirable government jobs. In Spain civil service competitions are public and typically involve the face-

[1] Fernández & Sosa (2005) and Castilla (2005) are notable exceptions.

to-face interaction of evaluators and job applicants. The reason underlying publicity is that Spanish civil service prides itself on being entirely merit-based and as such, presumably welcoming of popular scrutiny. A second important pillar of civil service recruitment's merit-based foundation is that applicants are selected by committees, it is never the case that a single evaluator is entrusted with the selection of applicants. This feature ensures that hiring does not rely on the idiosyncrasy a single individual.

This project focuses on a specific competition, Administradores Civiles del Estado (henceforth, ACE), which consists of four qualifying rounds of testing; applicants who succeed at round four are automatically hired and become permanent civil servants.

The first part of this research analyzes quantitative exam data gathered using open Internet sources. I use status characteristics theory to understand why women score higher than men in some exams (i.e. exam 2) but not in others (i.e. exam 3). In the ACE competition exams are highly similar but differ in two important aspects that are theory-relevant. The purpose of this section is twofold. First, some exams evaluate female-typed skills while others assess masculine or neutral abilities. I assess whether status characteristics theory adequately explains the actual competition outcome. Do women score lower in exams that evaluate masculine/neutral abilities? Do women enjoy an advantage relative to men when the skills to be assessed are female-typed?

Second, although all exams involve the direct interaction of judges and applicants, exams differ in the degree of structure of such interactions. In some exams applicant-judge interactions are rigid and minimal (i.e. exam 1 and 4) while in others applicants and judges engage

in a fairly natural dialogue (i.e. exam 3). Ridgeway (1997) has argued that mixed-sex interaction will prompt sex-categorization, which will in turn activate the use of gender status beliefs to guide attitudes and behavior (Ridgeway 1997). Other scholars have noted how ambiguity facilitates the influence of gendered expectations on evaluations (Heilman & Parks-Stamm 2007); vagueness create gaps that are susceptible to be filled with gendered subjectivity (Nieva & Gutek 1980) so that information fits a preconceived (thus preferred) outcome (Fiske & Taylor 1991).

Considering these various approaches, I argue that the degree of structure of judge-applicant interactions will impact the extent to which gender status beliefs will be acted upon by individuals in the setting. In other words, contexts where applicants and evaluators interact more freely (i.e. engage in an actual conversation) will disadvantage female applicants more than situations where interaction is direct but minimal or more structured. The proposed rationale is that interactive settings that are less-structured leave it to individuals' choosing whether or not to exercise behaviors based on their gendered expectations. Conversely, highly structured contexts serve to constrain or limit individuals' unconscious impulses to act on their gender beliefs. I argue (and provide empirical support) that classifying exams by their degree of structure is useful to understand the magnitude of SCT predictions.

The second part of this work builds on the main findings of part one and takes an in-depth look at what happens during the Q&A (Questions and Answers) portion of exam 3. I use double standards theory to evaluate whether specific behaviors of judges and applicants permit substantiating the experimental finding that men and women are assessed according to

different criteria even when they provide the same evidence of ability. I audio-taped and transcribed 83 Q&A sessions and hired two expert coders to obtain objective measures of the quality of applicants' answers and the difficulty of judges' questions.

Since judges' performance expectations are not directly observable, I measured behaviors (i.e. judges' interruptions) that reveal some of the cognitive processes taking place as judges evaluate applicants. Similarly, I measured various applicants' behaviors such as pauses and speech duration, which provide information about the quality and style of applicants' capacities.

In this work I try to take a fresh approach to the study of gender segregation by examining a novel setting with a magnifying lens. The main goal of this project is to help us gain an in-depth understanding of whether and how women are disadvantaged in the hiring process. Second, this work represents a two-fold methodological contribution: (i) first I use the method of conversation analysis to understand gender dynamics in a real world context; (ii) second, this is one of the first applications of experimentally established theory to a natural setting, as such this work illustrates how mechanisms found in controlled environments operate in much less structured contexts.

CHAPTER TWO
PRIOR RESEARCH

Gender segregation in employment refers to the unequal distribution of men and women across industries and jobs. Occupational sex segregation is a major source of labor market rigidity and economic inefficiency (Anker 1997). In addition, the asymmetrical distribution of male and female workers across occupations is associated with a broad range of workplace inequalities. First, occupational segregation is a major cause of the gender gap in wages and benefits. Second, female-typed jobs offer fewer promotion and on-the-job training opportunities (Farkas et al. 1991). Third, gender segregation in employment not only reflects hegemonic gender beliefs but also it contributes to perpetuate them (Ridgeway 1997).

Although the proportion of men and women in the labor force is approaching parity, survey data suggest that there is still substantial segregation across occupations, organizations, and industries (Anker 1997; Reskin 1993). Furthermore, recent studies claim that occupational segregation contributes substantially to the gender pay gap (Blau & Kahn 2006; Petersen & Morgan 1995). Since it is well-established that female workers earn less because they are more likely to fill positions that offer lower economic rewards, it is important to examine the processes that are preventing women from accessing the best jobs.

Supply Side & Demand Side Explanations

Explanations of the sources of occupational sex segregation are

typically categorized into two major approaches, namely supply and demand. Generally speaking, supply-side theories argue that men and women end up filling up different jobs as a result of their different preferences, natural abilities, rational investments, or biased self-assessments. In other words, since men and women are (or think they are) dissimilar, and confront different obligations, they are naturally, rationally, and culturally oriented toward different occupations.

For example, Becker's supply side argument is that men and women make differential investments in human capital in an effort to maximize income. Becker's human-capital theory argues that employees are rewarded for the value of the additional productivity brought about by their investments in skills. The author argues that the need for labor market specialization provides a strong incentive for a division of labor, which leads men and women to invest more in the areas where they each have a comparative advantage – labor market and household respectively. As a result, working women rationally save on labor-market effort by seeking jobs that require less human capital investment (Becker 1985).

A more sophisticated example of supply-side mechanism is one developed by Correll (2001, 2004) in an effort to move away from the overly simple gender socialization and rationality approaches. The author argues that widely shared cultural beliefs about gender and task competence differentially bias how men and women evaluate their own competence at career-relevant tasks. According to Correll this bias is the result of the internalization of a cultural belief about gender and a given skill into one's identity, or the expectation that other individuals will think this way (Correll 2001, 2004). In other words, existing beliefs about

gender lead women to use stricter standards to evaluate their own competence at male-typed abilities; these inaccurate self assessments will in turn influence women's career orientations and aspirations, and ultimately the jobs they pursue.

Demand side explanations contend that factors beyond employees can also lead to gender segregation. These factors include but are not limited to discrimination and rational employers' decisions[2]. However, the remainder of this chapter will focus on empirical evidence supporting the discrimination argument, which in some ways challenges rationality-type explanations.

Gender segregation is, at least partially, the result of aggregate individual gender-based judgments (Perry, Davis-Blake & Kulik 1994). Thus, some demand-side approaches to segregation focus on the role of employers' judgments in hiring and promotion. The economic perspective focuses on rationality and tries to explain why it is advantageous for employers to make the choices they make when recruiting and promoting job applicants – e.g. statistical discrimination theory by Phelps (1972) and Arrow (1973). Conversely, the status based perspective draws on the notion that human behavior and decisions are not merely rational nor agentic. While economic theories such as statistical discrimination assume the source of bias leading employers to prefer one group over another is informational, status discrimination theories assume the source of bias is cognitive. In statistical discrimination models, employers are perfectly rational and maximize expected utility. Bias enters hiring and wage

[2] For example, Fernández and Sosa (2005) evaluate the extent to which network-based hiring practices (i.e. referrals) affect the demand for female and male workers.

decisions through external constraints (i.e. lack of information) but is not inherent in them. In contrast, status discrimination theories assume that actors' cognitive abilities are limited (see Correll & Benard 2006).

Stereotype & Status-Based Discrimination

Scholars have argued that it is the unequal treatment and evaluation of men and women by employers that produces biased hiring decisions resulting in the patterns of gender segregation observed at a macro-level. According to this view, it is not the case that men and women are so different; rather individuals are perceived and assessed differently based on their ascribed characteristics.

Gender stereotypes can be about what men and women are (i.e. descriptive) or about what men and women should be (i.e. prescriptive). Both properties affect how women are evaluated and treated in career-relevant settings (see Heilman & Parks-Stamm 2007). Specifically, stereotypes disadvantage women in employment contexts by hindering their efforts to achieve status in the workplace. Women are often caught up in a double bind since displaying the competencies required for top level jobs is culturally framed as incompatible with femininity (Heilman 2001; Heilman & Parks-Stamm 2007).

Descriptive stereotypes about women create the expectation that women are unlikely to be successful at male-typed jobs; these performance expectations influence the way information about individuals is processed. Specifically, expectations affect what information about an individual is attended to, how it is interpreted, and what it is recalled when evaluations and decisions are made in the workplace. Thus,

expectations can create gender bias in evaluative judgments at the point of hiring and subsequent career relevant situations (Heilman 1995, 2001). Expectations promote the view that women are unfit for a job, and as such unlikely to succeed at it. Thus, stereotype-based performance expectations directly affect women's chances of being recruited or promoted into male-typed positions.

Gender status theories argue that discrimination arises because generalized beliefs about the relative performance capacity of men and women influence evaluations of workers. Although these are not theories of labor market discrimination per se, SCT and DST make predictions about when and how women will be discriminated against in hiring, salary, and promotion decisions. The key notion in status-based discrimination is that employers' assessment of future employees will be shaped by shared beliefs determining that a category of the social distinction (e.g. for gender, males) has more value than the other (e.g. female). Thus, employers will implicitly anticipate superior performances from male job applicants than for their female counterparts and subsequent evaluations will be biased in favor of men.

Research has empirically supported the existence of gender bias in employee selection processes (see Davidson & Burke 2000) with male applicants generally recommended for hire and seen as more likely to succeed than female applicants with identical credentials. Studies have shown that despite producing identical work, a woman's work is often regarded as inferior. Research in organizational psychology found that unless the quality of the work product is incontrovertible, women's accomplishments are undervalued (Heilman 1995).

Various studies have proved that group members often hold lower performance expectations for women than men (Berger, Rosenholtz, & Zelditch 1980; Lockheed & Hall 1976; Meeker & Weitzel-O'Neil 1977) and give women fewer opportunities to participate than men (Meeker & Weitzel-O'Neil 1985; Ridgeway & Berger 1986). Similarly, other works demonstrate that equally competent performance by men and women is perceived as more indicative of skill and ability in men than in women (Deaux & Emswiller 1974; Foschi, Lai, & Siegerson 1994). For example, Deaux and Emswiller (1974) showed that a successful performance is not treated as very informative of a woman's competence; rather, her success is explained away by factors unrelated to her capacities (i.e. luck).

The use of different standards to evaluate men and women's competence has been confirmed in a variety of settings (see Foschi 2000). For instance, in an experiment Foschi and colleagues (1994) recreated features of a hiring decision that involved the examination of files of fictitious applicants for professional jobs. Subjects had to make a recommendation about hiring the fictitious male or female applicant. Although female participants did not display the use of double standards, the results from male participants indicated otherwise. For male subjects, fictitious male applicants were preferred when they were slightly more qualified, but fictitious female applicants did not enjoy the same advantage when they had slightly better qualifications (Foschi, Lai, & Siegerson 1994). Other studies suggest that both men and women rate the quality of men's work higher than that of women's work when they are aware of the sex of the person to be evaluated, but not when the person's gender is unknown (O'Leary & Wallston 1982).

In an audit study, Steinpreis and colleagues (1999) examined whether faculty would be influenced by the gender of the name on a CV in determining hireability and tenurability. Fictitious CVs were submitted to real academics. Both male and female faculty were significantly more likely to hire a potential male colleague than an equally qualified potential female colleague. In addition, both male and female faculty were more likely to positively evaluate the research, teaching, and service contributions of male job applicants than that of female job applicants with identical records (Steinpreis, Anders, & Ritzke 1999). Interestingly, participants in Steinpreis study were four times as likely to write down cautionary comments when reviewing the CV of a fictitious female tenure candidate than when reviewing CVs of fictitious males. Comments included notes such as "we would have to see her job-talk", "it is impossible to see such a judgment without teaching evaluations", or "I would have to see evidence that she had gotten these grants and publications on her own".

Using a similar methodology, Neumark (1996) conducted a study in which male and female fictitious job seekers were given similar CVs and were sent to apply for jobs waiting on tables at the same group of restaurants. In top restaurants, the female applicant's probability of getting a job offer was 50% below that of the male (Neumark 1996).

Another study examined the impact of the adoption of blind auditions by symphony orchestras where a screen is used to hide the identity of the performer (Goldin & Rouse 2000). The authors confirmed that the screen increased the probability that a woman would be selected.

The switch to blind auditions in 1996 accounted for 25% of the increase in the percentage female in the top five symphony orchestras in the US.

Negative evaluations in selection processes have been found to occur particularly for male-typed jobs (Davison & Burke 2000). There are similar findings in investigations concerning competence assessments and performance evaluations. For example, a study on performance evaluations in a large multinational financial services company demonstrated that women were rated less favorably than men in line jobs but not in staff jobs (Lyness & Heilman 2006).

Thomas-Hunt and Phillips (2004) found that women (who were equally qualified as men) were perceived by others as less expert than men, were less influential, and felt less confident about their impact on the group. Their results support the notion that the possession of expert knowledge is likely to be more beneficial for men than for women. The authors' findings suggest that women are penalized when they possess the same expertise as men (Thomas-Hunt & Phillips 2004).

As can be gathered from the discussion above most studies of gender bias in the selection process are done in controlled or semi-controlled settings. These studies can be viewed as alternatives to more traditional approaches to the study of discrimination. For example, a classic approach that examines discrimination indirectly is one that analyzes the sources of the gender wage gap by trying to account for as many as possible productivity-related characteristics for men and women. The pay gap is statistically decomposed into two components: one due to gender differences in measured characteristics, and the other unexplained and presumably due to discrimination. But any approach that relies on a

statistical residual is open to the question of whether all the relevant explanatory variables were included in the model (Heilman & Parks-Stamm 2007). An additional problem that other authors have noted is that studies using data on individuals who already have jobs cannot adequately identify mechanisms such as discrimination which occur prior to getting hired (Fernández & Sosa 2005). By contrast, social psychological experiments have made important contributions to explaining why employers might prefer men over women to fill the best jobs. Although cognitive approaches do not directly analyze occupational segregation, they examine processes that have clear implications for it.

CHAPTER THREE
THEORY

This work relies on various theories of the expectation states research program, an on going line of analysis that explains how structures of inequality emerge, are maintained, and translate into material and nonmaterial advantages for certain social groups (Correll & Ridgeway 2003). Expectation states theory offers an all-inclusive explanation to situations felt and described by socially devalued groups – i.e. feeling ignored or overridden in discussions with others. These small inequalities that emerge in interaction have important cumulative effects. Thus, it is crucial to understand how preconceived expectations for high and low status group members develop and influence social interaction.

Expectation states theory explains the formation of status hierarchies in contexts where individuals are compelled to solve a problem or achieve a goal in a group. In other words, the theory holds in collective and task oriented situations. A key notion is that such situations make it necessary and useful for actors to take into account other group members' contributions, which in turn compels actors to predict and weight the relative quality of others' suggestions. The setting where I apply this theory and its extensions meets these scope conditions in several respects. First, exam judges' are entrusted with the task of selecting the best performing applicants; thus, evaluators will consider how each applicant measures against other applicants and/or against their abstract notion of competent or deserving applicant. Thus, judges are motivated to anticipate applicants performances so as to assist and

economize their decision-making process. Applicants are similarly task-oriented insofar as they are aware job spots are limited and only the best performers will be selected. Thus applicants are in direct competition with all other examinees and, understandably, are highly compelled to demonstrate their superiority relative to others. Furthermore, as will be explained later in this chapter, EST and its different branches have, in more recent developments, relaxed some of the specifications under which EST predictions are argued to be correct leaving no doubt that the selected setting meets the theory's scope conditions.

Expectation states theory argues that individuals form expectations about the relative competence of group members to contribute to a shared goal. When performance expectations for a group member are high, the person will enjoy a series of privileges in social interaction (Berger et al. 1972). Performance expectations are unconscious anticipations that are shaped by a variety of factors such as (1) socially significant or status characteristics, (2) social rewards, and (3) behavioral patterns. I will discuss these factors in detail in the following paragraphs.

Status Characteristics Theory

High or low expectations largely originate in attributes such as race or gender for which there are broadly shared cultural beliefs. More formally, status characteristics are categorical distinctions among people; different categories (e.g. for gender, male and female) have attached to them hegemonic beliefs associating greater status and competence to one category (i.e. men) of the distinction and not the other/s (i.e. women) (Berger et al. 1977, 1972).

Status characteristics theory describes how socially meaningful distinctions lead to inequalities in rates of participation, influence on others, and evaluations of task competence. According to SCT, actors implicitly expect superior performances from those with the more valued state of a characteristic (e.g. men) relative to those with the less valued state (e.g. women). Performance expectations have a self fulfilling component; they bias information processing as means of maintaining themselves. Thus, performance expectations affect what evidence is attended to and recalled, and how that evidence is interpreted when making decisions about individuals. Since high status actors are expected to offer more competent performances, they receive more opportunities to make contributions, have more influence on others and have their performances evaluated more positively (Correll & Ridgeway 2003).

To understand how generalized cultural beliefs function, it is useful to contrast them with stereotypes. Separating individuals into differentiated social categories encourages preference toward one's own group (Brewer & Brown 1998). Conversely, status beliefs are "social representations that consensually evaluate one category as more status worthy and competent than the other" (Correll & Ridgeway 2003, p. 32). This means that even though socially devalued groups such as women may favor their own group (i.e. other women), they will also be aware of and accept, or at least concede, men's superior social status (Ridgeway & Erickson 2000). Thus, gender status beliefs reflect a cultural system representing what we think most people accept as true about men and women (Deaux & Kite 1987). Because status beliefs function as schemas (Ridgeway 1997), even those who do not personally endorse their content

(i.e. many individuals may disagree that men are better than women at math) are likely to be aware of their existence and thus have their judgment and behavior influenced by them (Foschi 1996; Lovaglia et al. 1998; Ridgeway & Correll 2004; Steele 1997). While the specific content of gender beliefs can differ across contexts and cultures, their status component ensures that greater value will be attached to the superior category (i.e. men) but not the other/s (Conway et al. 1996).

There are five assumptions that connect status beliefs to behavior (Balkwell, 1991) and these are (a) salience, (b) burden of proof, (c) sequencing, (d) aggregation, and (e) observable behavior. Next, I turn to a discussion of these five tenets.

The salience premise states that for any attribute (e.g. class, gender, race) to impact performance expectations, the attribute must be important for actors in the setting. A status characteristic becomes important or salient when it differentiates individuals in a given context or when the characteristic is perceived to be related to the task. For instance, gender is salient whenever men and women interact. Gender would also be salient in a context where a group of women work on a task requiring verbal ability, a stereotypically female skill. Thus, specific contexts and their social composition shape how status characteristics affect performance expectations. The same characteristic (e.g. fluency in a foreign language) can be an advantage to an actor in one setting (e.g. a group of monolingual speakers), have no impact in another (e.g. a group of bilingual people), and be a disadvantage in a third context (e.g. a group where all members are native speakers). Importantly, this implies that no status characteristic advantages or disadvantages an actor in all settings.

Salient status characteristics have been shown to impact attitudes and behavior in collective and goal oriented situations. Later advances of the theory have demonstrated that status characteristics shape behavior in a broader range of social contexts than originally specified by the theory's scope conditions. In fact, status characteristics have been shown to matter in all contexts where actors are compelled to anticipate their own behavior relative to others and/or the behavior of others (Lovaglia et al. 1998; Steele 1997; Foschi, Lai & Sigerson 1994; Correll 2004). Researchers have relaxed the collective orientation condition because the logic of the theory only requires that some feature of the setting encourages actors to predict the relative value of their own or others' contributions. For example, pressure to measure oneself against abstract others can also appear when actors are in individual evaluative settings (e.g. Correll 2004; Erickson 1998; Lovaglia et al. 1998; Steele 1997) and, when actors are evaluators but not performers (Foschi, Lai & Sigerson 1994). In terms of the Spanish civil service exams, evaluators may not be collectively oriented when assessing applicants, but they certainly are compelled to anticipate prospective applicants' relative performances. Since the civil service applicant pool is heterogeneous in terms of gender, status beliefs will encourage the use of different ability standards for male and female applicants; this in turn will lead to higher evaluations of men.

Second, the burden of proof assumption tells us that actors' default response is to assume salient status characteristics are relevant to the task at hand. The challenge consists on showing that a salient status characteristic is irrelevant and should not be taken into account when anticipating others' performances. In other words, something in the

setting needs to explicitly dissociate the status characteristic from the task; otherwise, individuals will behave as if the characteristic matters. For these reasons, diffuse status characteristics such as gender have limited but omnipresent effects across a multiple contexts even when they are not directly related to the task.

Third, the sequencing assumption specifies that performance expectations formed in one encounter carry over to the next, even if the individuals in the encounter change. For example, if a man interacts with a woman who displays greater task competence than he does, this encounter can positively impact the performance expectations the man forms for women in future interactions (Pugh & Wahrman 1983). For these reasons, gender scholars have identified interactional settings as contexts where inequalities can be created but also where social change can occur (Ridgeway & Correll 2000).

Fourth, individuals in groups typically differ in more than one status characteristic, the aggregation assumption examines how the information associated with multiple characteristics is combined to form aggregated performance expectations. Multiple status characteristics often generate inconsistent expectations for performance – e.g. male nurse. SCT offers a method for incorporating all salient status information to determine the performance expectations group members will form. A nice feature of my civil service context is that individuals do not generally differ in other important status attributes such as nationality, ethnicity, age and so forth. If this were not the case, it would be difficult to argue that gender alone is causing the observed outcomes. In this sense, the setting I examine, although not controlled in a strict sense, is highly structured and

additionally has the important advantage of being a natural context with real consequences for actors.

The fifth assumption describes how performance expectations which are unobservable translate into tangible behaviors. The higher the performance expectations of one actor over another, the more likely the first actor will be to receive opportunities to participate and take such opportunities. Also, higher status actors will have their performances evaluated more positively and will command more influence over others. In my setting, I argue that evaluators' expectations will be revealed by a series of behaviors; it is useful to view the behaviors I selected and measured as specific illustrations of the more general behavioral predictions above. For example, measuring when judges interrupt applicants constitutes a denial of an opportunity to participate for applicants. In fact, since interruptions are negative sanctions administered in front of other evaluators their effects may also impact other judges' performance expectations.

At the beginning of this chapter I noted that, in addition to status attributes, reward distributions can have independent effects in shaping the performance expectations of actors in a setting. Several scholars have shown that when group members are given differential rewards, they use the reward differences to infer ability differences. For example, Stewart and Moore (1992) showed that allocating differential pay levels to participants in an experiment generated influence structures among them during interaction. These results highlight how the unequal possession of valued goods generates status distinctions. In the civil service exam setting that will be described in thoroughly in Chapter 4, judges do not

administer rewards but sanctions (i.e. interruptions). It seems reasonable to assume that penalties work in the same manner. Succinctly, applicants who receive more interruptions may come to be viewed as deserving such penalties. Thus, other judges may infer lower ability from the unequal distribution of interruptions.

Finally, a third factor that can have independent effects on performance expectations is the behavioral patterns that develop among two or more actors. A variety of assertive verbal and nonverbal cues (e.g. sitting at the head of the table, having an upright, relaxed posture, speaking up with a confident tone, maintaining eye contact) have been shown to make a person's ideas sound better and increase influence (see Dovidio & Ellyson 1985; Ridgeway 1987). In diverse groups, actors' status characteristics determine behavioral interchange patterns (Smith-Lovin & Brody 1989). This is relevant to my research because I measured and examined behaviors such as interruptions, pauses, and speech duration. Interrupting is a leader-type behavior and thus will be enacted more often by men and directed usually at women.

Double Standards Theory

Double standards theory (DST) extends SCT to propose that status characteristics also affect the standards individuals use to determine if a given performance is indicative of ability (Foschi 1989). There are two processes involved in the use of double standards. First, status characteristics distinguish actors in the setting – i.e. a mixed-sex setting. Second, the available evidence indicates that individuals from both groups possess a given attribute to similar extents. Double standards are the

practice of using different criteria to interpret similar evidence.

DST argues that stricter standards are applied to members of devalued groups and, as a result, conclusions about the extent to which they posses the attribute based on the available proof are reinterpreted through a sort of lens. A strict standard for ability requires more evidence of competence than does a lenient standard. Conversely, a strict standard for lack of ability tolerates less evidence of incompetence than does a soft standard (Biernat & Kobrynowicz 1997; Foschi 1989).

When high status individuals (e.g. men) perform well, evaluators' expectations are met and there is no cognitive dissonance that might need resolving. When low status individuals (e.g. women) perform well, the opposite happens; good performance appears inconsistent with what was anticipated. Thus, evaluators' will tend to doubt women's competence and will further scrutinize it (Foschi, 1996; Foschi, Lai & Sigerson 1994).

The notes at the margins of the fictitious female CVs in Steinpreis and colleagues' study (1999) (see Chapter 2) may not provide sufficient data for a standard statistical test. Nonetheless, these cautionary comments mentioned by the author reflect a qualitatively important finding, namely that identical evidence provided by male and female applicants is interpreted based on different criteria. Essentially, information contained in women's CVs is not fully trusted; the performances of low status actors—even when objectively equal to that of their high status counterparts—are less likely to be judged as demonstrating competence. Empirical evidence supports these predictions for gender in contexts where individuals evaluate others and when they evaluate themselves (Foschi 1996; Correll 2001, 2004).

Outside the expectation states framework, the research by Biernat and colleagues on stereotypes and shifting standards is also relevant (see Biernat & Fuegen 2001; Biernat & Kobrynowicz 1997). The core idea of their work is that standards change as a function of who is being evaluated but also that different conditions result in either a more lenient or a stricter standard for low status groups (Biernat & Kobrynowicz 1997; Biernat & Fuegen 2001). For example, the authors argue that a more lenient standard of ability for women would result if a specific woman is compared against other women on a given dimension (i.e. within category comparisons). According to Biernat and associates, it is often the case that individuals have lower minimum standards for women and other devalued groups, and higher confirmatory standards for men and high status groups in general. Biernat and Fuegen (2001) found that in a simulated hiring context women were more likely to be short listed in the selection process but that male applicants were more likely to be hired (Biernat & Fuegen 2001). The authors conclude that lower minimum standards for women make it easier for them to pass an initial screening process but that higher confirmatory standards make it more difficult for them to pass the scrutiny required to be hired (Biernat & Fuegen 2001). I will discuss the implications of the shifting standard model in my context when I provide related concrete hypotheses (Chapter 5).

When gender is salient in goal oriented settings, SCT and DST predict that men will have an advantage over women because they will be expected to perform better. Provided that a context is not female typed, higher performance expectations for men lead to three main theoretical predictions: (1) men will be given more opportunities to participate or

make a contribution, (2) men will have their mistakes judged by more lenient standards; stricter standards will be used for women, and (3) men will have their performances evaluated more positively than women. Experiments confirm that a number of status characteristics, including race, gender, and level of education systematically organize the appearance of competence, influence and deference (Lovaglia et al. 1998; Ridgeway 2001).

Degree of Structure in Interaction

Although a setting needs not involve direct interaction for gender status theory predictions to operate, Ridgeway (1997) has argued that relating to a concrete other is sufficient to trigger gender status processes. Ridgeway (1997) understands social interaction as a complex phenomenon that requires to be simplified before it can be coordinated. Simplification begins to occur when individuals develop a minimal definition of who "self" and "other" are in a given context; preliminary definitions are reached by contrasting self and other on dimensions where similarities and differences are perceived to exist. Empirical evidence demonstrates that sex serves as a primary categorization system in Western society (Fiske 1992) and that individuals automatically and unconsciously sex categorize any specific other to whom they relate (Brewer & Lui 1989; Stangor et al. 1992). Subsequent categorizations such as occupational roles become nested in gender (Brewer & Lui 1989), taking on slightly different meanings as a result. Most importantly, sex categorization prompts the use of gender stereotypes (including status beliefs) to guide attitudes and behavior (Blair & Banaji 1996).

Heilman and Parks-Stamm (2007) have argued that "the more ambiguity there is in a given context, the more inference is required for evaluation, and the less guidance there is about the correct outcome of an evaluation" (Heilman & Parks-Stamm 2007, p. 54). The authors' argument is that ambiguity creates gaps that can be filled with gender biased subjectivity (Nieva & Gutek 1980) so that information fits a preferred outcome (Fiske & Taylor 1991).

From the concepts briefly outlined above, I make the following elaboration: contexts involving less structured interactions will encourage actors to use gender status beliefs to a greater extent. Thus, while SCT predictions will work regardless of the level of structure in judge-applicant interactions, I argue that the magnitude of these predictions will be greater when actors interact more naturally and freely than when they do so in more rigid or scripted ways[3].

Other scholars such as Mueller and colleagues have attempted to better define interaction and have proposed to operationalize the concept as a continuous measure (see Mueller, Mulinge, & Glass 2002). The authors understanding of degree of interaction (i.e. as frequency or rate; see Mueller, Mulinge, & Glass 2002) is quite distant from the approach I take in this work. I have argued that although all testing rounds involve the face to face interaction of evaluators and applicants, exams differ in the degree of structure of applicant-judge interactions. Degree does not refer to duration, frequency, or rate; rather a greater or lesser degree of interactional structure is accorded as a function of qualitative aspects

[3] Ridgeway (1997) makes this point but mine is a more explicit discussion and will provide empirical support.

characterizing the exchange. Degree refers to the level of structure (or lack of structure) of applicant-judge conversations. A highly structured interaction is one governed by explicit rules. In these situations actors' behaviors are limited by a series of norms which presumably substitute and prevent the use of more personal criteria on which to base their behaviors. Since individuals' attitudes and actions are often shaped by preconceptions based on status beliefs and stereotypes, it could be argued that low status actors benefit when the rules of interaction are more explicit. Conversely, in other more informal types of face-to-face encounters actors' behaviors are less constrained and as such, they are more likely to reflect actors' assumptions and prejudices. In less structured or more interactional contexts, low status actors are often at a disadvantage because there will be more opportunities for others to treat them according to the performance expectations they hold for them.

I argue that in less structured interactions, allowed and prohibited behaviors are not clearly defined. Thus, it is up to the individual to fill this vacuum, and actors will exercise such freedom in a gender biased manner. Thus, if evaluators have implicit preferences for male applicants, a less structured setting will enable evaluators to behave in ways consistent with these beliefs. In other words actors' status information combined with structural characteristics of the setting (i.e. more or less degree of structure in interactions) means more biased behaviors are likely to surface. For example, in a formal presentation audience members may think the speaker is not doing a good job and may feel compelled to correct or interrupt. But it is possible that the rules of the setting prevent audience members from doing so – i.e. questions or suggestions may only

be asked at the end. Conversely, contexts where interaction is less structured like an informal presentation may lack concrete rules. In the absence of explicit protocol, audience members will use their own criteria to orient their behavior. The gender and status literature suggests that audience members will be more forgiving with a male speaker (i.e. will not interrupt, fidget, display impatience) than with a female speaker, even when the two are equally competent/incompetent.

Because mine is a novel setting and as such, unfamiliar to most readers, in the next chapter (i.e. Chapter 4) I will provide a thorough description of the context of the study. Chapters 5 and 7 will offer specific hypotheses for the two analytical parts of this work.

CHAPTER FOUR
CONTEXT

This research relies on theories which rest on the notion that widely held beliefs about gender exist and are used constantly to organize social relations. The specific content of gender beliefs is not presumed to be constant across cultures[4]; nonetheless, a key aspect of these theories is that greater status is always associated only with a category of a social distinction and not the others. For instance, in Spain men may not be perceived to be naturally good at math like it is the case in the US[5] but rather, at negotiation. While the specific content of beliefs is different (i.e. mathematical vs. negotiation skills), both aptitudes are socially desirable. What gender status theories argue is that high status actors (i.e. men) will be perceived to possess socially valued abilities (whatever these are) to a greater extent than low status actors (i.e. women). This means that if a new ability becomes socially valued, individuals will be inclined to associate such ability to high status groups such as men, whites, the educated and so forth.

In this section I will justify the claim that men enjoy greater social status in Spain like in many other societies. First, I will examine empirical evidence substantiating that gender-role attitudes in Spain are similar to those existing in the United States and Europe. Second, I will offer a

---

[4] Nonetheless research has demonstrated that gender stereotypes are very consistent across time and cultures (Williams & Best 1990) and that they are pervasive, widely shared, and very resistant to change (Dodge, Gilroy & Fenzel 1995).

[5] See Correll 2001 for a review of the cultural association men-math ability and the absence of actual differences between men and women's mathematical ability in the United States.

description of the context and specific setting of this study. This description will be extended in chapters five and seven, which will specify hypotheses for the two distinct analytical sections of this work.

Gender Beliefs in Comparative Perspective

Chapter 3 discussed that hegemonic gender beliefs accord men greater status worthiness than women, and explained how individuals use these cultural schemas constantly in their assessment of social situations and subsequent behavior. Thus, both status characteristics and double standards theory elucidate how shared and deeply rooted beliefs about gender translate into tangible hurdles and disadvantages for devalued groups in areas as crucial as education and employment.

In the following paragraphs, I will examine and discuss empirical evidence suggesting that gender-role attitudes in Spain resemble those that exist in the United States and elsewhere in Europe[6]. I analyzed data from the International Social Survey Programme (ISSP), which is a collaboration among a total of 41 nations which conduct harmonized surveys about topics of ample interest for social science research. I examined a set of attitudinal survey items of the 2002 module on gender attitudes; this analysis will help support the notion that men are higher status relative to women in Spain. My general argument is that in Spanish society, like in many others, men are considered to be better than women at the things that matter (i.e. men are higher status than women).

[6] Europe includes EU15 member states (except Greece, Belgium, and Luxemburg, which were not surveyed for gender attitudes in 2002), Norway, and Switzerland.

A total of 2,471 respondents from Spain and 19,309 from Europe and the US combined participated in the ISSP module on gender attitudes in 2002. Respondents were asked to rate a series of statements on 5-point scales where 1 was "strongly agree" and 5 was "strongly disagree". Below I present some descriptive results concerning respondents' attitudes toward men and women's perceived roles and preferences.

Table 1 ISSP Survey Results on Gender Attitudes in Spain and the United States & Europe, 2002

Strongly Agree/Agree That…		**Spain**	**US/Europe**
Children suffer if mother works		52%[1]	43%
	N	1,253	7,984
	Mean[2]	2.84	3.00
Family life suffers if women work		54%	41%
	N	1,331	7,711
	Mean	2.75	3.05
What women really want is home/children		42	34
	N	984	6,073
	Mean	3.07	3.20
Women's place is the household		24	19
	N	601	2,977
	Mean	3.61	3.72

[1] Percentage of respondents who answered "strongly agree" or "agree".
[2] Where 1= strongly agree and 5= strongly disagree.

About 32% (N=785) of Spaniards believe that working mothers cannot have a warm relationship with their children. About 21% (N=3,964) of respondents shared this view in other European countries and the US combined. Similarly, approximately 52% (N=1,253) of

Spanish respondents agreed or strongly agreed that children suffer if their mother works outside the home. About 43% (N=7,984) of respondents in Europe and the US answered likewise. In the same vein, 54% (N=1,331) of respondents in Spain agree or strongly agree that family life suffers if women work outside the home. About 41% (N=7,711) of respondents in Europe and the US answered the same.

Regarding the perception of women's preferences and aspirations, 42% (N=991) of Spanish respondents agreed that what women "really want" is to stay home and take care of their children. About 34% (N=6,073) of US and European respondents shared the same views. Roughly 24% (N=601) of Spaniards agreed that men's job is outside the home and women's job is in the household. In the US and Europe combined, 19% (N=2,977) of respondents answered likewise.

Table 1 summarizes the results discussed. Percentages represent the proportion of respondents who "strongly agreed" or "agreed" with the statements on the left hand column (i.e. "Children suffer if mother works"). Table 1 also shows the mean values for both Spanish and US/European respondents. Higher values indicate less agreement with the statements on the left (i.e. 1="strongly agree" and 5="strongly disagree").

From this examination of gender role attitudes, it seems that generalized beliefs exist in Spain about men's superior social status relative to women. To the extent that civil service evaluators are aware that gender beliefs exist in the culture, they will subconsciously draw from to them to orient their attitudes and behavior even if they do not personally endorse their content. The figures above also suggest that, on average, Spanish respondents are somewhat more traditional with regards

to gender-role attitudes than European and American respondents. Thus, there are reasons to believe that gender operates as a status characteristic in Spain, with men being seen as higher status than women. At the same time, if civil service evaluators hold views that are, on average, comparable to the rest of the Spanish population, the implication could be that the setting only permits a conservative evaluation of the theory. In other words, the study's findings may be more extreme in Spain than in other countries simply by virtue of the greater traditionalism.

But the literature suggests that the more public scrutiny to which an evaluator is subjected, the more likely hiring is to be egalitarian with respect to ascribed characteristics. Prior research demonstrates that evaluators are motivated to be more accurate they are held accountable for their decisions (Simonson & Nye 1992). Accountability can weaken the use of expectations by encouraging more effort in information search and less superficial processing of information. Foschi (1996) found this effect when experimental subjects felt they would be held accountable for their assessments (Foschi 1996, p. 251). In addition, accountability may motivate evaluators to present themselves in favorable terms (Klimoski & Inks 1990). Taken together, these arguments suggest that civil service evaluators are more likely to be attentive to issues of equity and fairness due to the greater public scrutiny to which recruiters are held. The reason why these exams are public is precisely to facilitate the kind of popular scrutiny that would challenge an erroneous decision. In other words, this selection process designed this way precisely to avoid bias and favoritism.

The Spanish Public Sector

The first challenge for researchers interested in the study of hiring practices is to gain access to actual data. Virtually all contexts where job applicants are evaluated are restricted to outsiders. This work takes advantage of a unique exception, and examines a novel hiring setting where access is permitted making it feasible to observe and collect data on real evaluations and hiring decisions made in the course of direct interaction. In Spain, exams to become a government employee are public and consist of a series of testing rounds that usually involve the live interaction of evaluators and applicants. The this setting is public is that Spanish civil service prides itself on being entirely merit-based; as such, the process is completely transparent and open to popular scrutiny – status beliefs should therefore not operate here. A second and related important pillar of civil service recruitment's merit-based foundations is that applicants are selected by committees, it is never the case that a single evaluator is entrusted with the selection of applicants. These characteristics guarantee that hiring does not rely on the whims and idiosyncrasies of single individuals.

Although some jobs in the Spanish civil service are predominantly filled by women, the best positions in Spain's public sector are still largely occupied by men. Civil service positions fall into four major categories, namely groups A, B, C, and D[7]. Group A jobs are the best

[7] Some examples of group A jobs and their ISCO equivalent are: Abogados del Estado (1110 Legislators), Diplomáticos (1120 Diplomatic Representative), Ingenieros Navales (2145 Naval Engineer), and Médicos Titulares (2221 Medical Doctors). Group B jobs are, for instance, Diplomados en Estadística del Estado (3434 Statistical and Mathematical Professionals), and Ingenieros Técnicos Aeronáuticos (3115 Engineering/Aeronautics Technicians). As for group C: Técnicos Auxiliares de

paying and most prestigious, while positions in groups C and D require fewer years of education but also pay less well.

Table 2 Percentage of Women in Group A
Civil Service in Spain by Year

Year	% Women
1996	29.5
1997	29.7
1998	30.4
1999	30.9
2000	31.9
2001	32.4
2002	33.4
2003	34.1

Source: Mujeres en Cifras 2003

As Table 2 shows, in 2003 women filled about 34% of group A jobs. Even though these figures may appear optimistic and the general pattern shows an increase of women since 1996, it is important to note that female applicants make up between 60% and 70% of the initial pool in many group A competitions, and around 50% in others[8] (See Table 3). If we assume that skill is distributed similarly among applicants, and

Informática (3121 Computer Assistants), and Técnicos Especialistas en Reproducción Cartográfica (3118 Draughtsperson, Cartographical). Finally an example of group D jobs is General Auxiliar de la Administración del Estado (4212 Tellers and other Counter Clerks).

[8] Some exceptions are Diplomats and Legislators where women make up about 40% of the initial applicant pool.

evaluations are objective, the number of women that should have been recruited should almost double those in Table 2.

Further inequalities can be detected if Group A aggregate numbers are broken down by pay level. There are four levels within group A – higher levels (e.g. level 29) correspond to greater salaries. As Table 4 shows the higher the salary or pay level the smaller the proportion of women.

Table 3 Percentage of Women in Initial Applicant Pool in Various Group A Competitions

	% Female
Abogado del Estado (Legislator)	42%
Cuerpo Diplomático (Diplomats)	41%
Inspector de Hacienda	47%
Inspector Fiscal	69%
Juez/Notario (Judge/Notary)	51%
Registrador	51%
Secretario Judicial	72%
Administradores Civiles	64%

Source: Bagüés 2005[9]

Table 5 illustrates that there have been a small but constant upward increase in the proportion of women in all pay levels since 2002. However this increase has varied considerably in magnitude.

[9] Bagüés, M. 2005. ¿Qué Determina el Éxito en unas Oposiciones? FEDEA Working Paper 2005-01.

Table 4 Percentage of Women by Pay Level in Group A Civil Service in Spain 2002-2007

	2002	2003	2004	2005	2006	2007
29 (High)	22.07	23.33	23.99	25.04	27.58	28.75
28	29.17	30.24	32.33	34.21	35.34	36.63
27	34.34	35.88	35.96	36.80	37.06	38.59
26 (Low)	41.68	42.46	43.30	44.33	45.74	46.63

Source: Instituto de la Mujer. Data: Boletín Estadístico del Personal al Servicio de Administraciones Públicas - Registro Central de Personal.
Available at: http://www.mtas.es/MUJER/mujeres/cifras/

Table 5 Annual Increase of Women in Group A Relative to Previous Year by Pay Level (2002-2007)

Year Interval	Level 29	Level 28	Level 27	Level 26
2002-2003	1.26	1.07	1.54	0.78
2003-2004	0.66	2.09	0.08	0.84
2004-2005	1.05	1.88	0.84	1.03
2005-2006	2.54	1.13	0.26	1.41
2006-2007	1.17	1.29	1.53	0.89

Figure 1 below shows that increases in all pay levels fluctuate considerably from one year to the next. For example, the Level 28 line shows a 1% increase and a 2% increase in the proportion of women from 2002 to 2003 and 2003 to 2004 respectively. However, after 2004 increments slow down considerably.

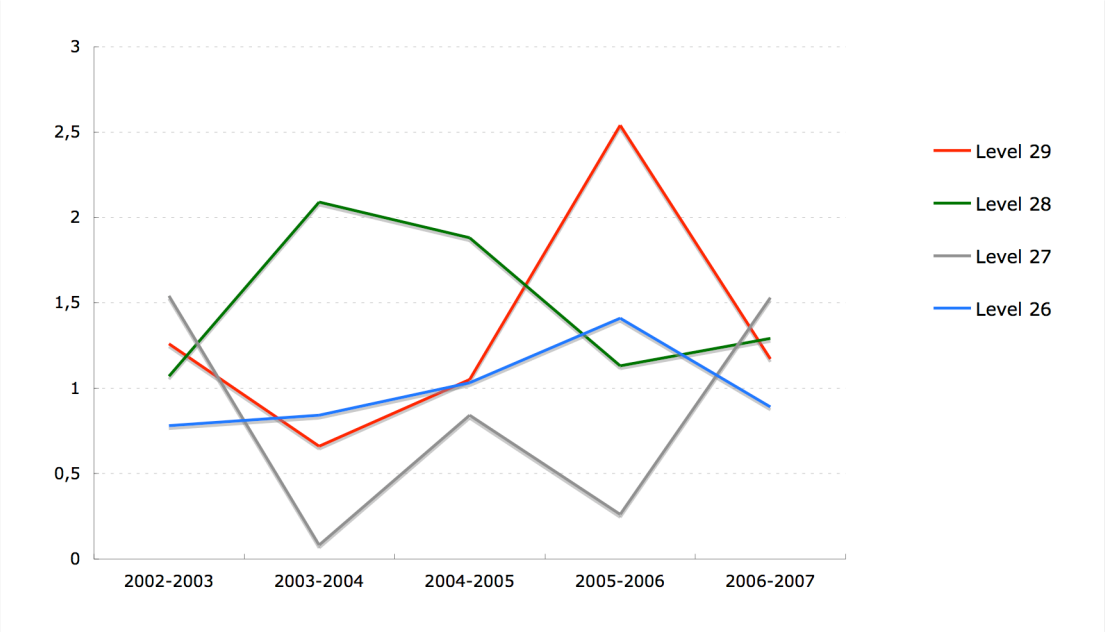

Figure 1 Percentage Increase of Women by Pay Level & Year

In the discussion above, three ideas are important and should be highlighted. First, women are underrepresented in the best positions of the Spanish public sector. Second, change is happening but at rather slow and fluctuating rates. Third, women are highly interested in attaining these positions thus, the system may not be opening up at the speed that women might deserve.

In Spain the public sector means above anything else employment stability and exceptional work conditions. According to a report by ANECA[10] about 45% of university students in Spain report they intend pursue a career in government upon graduation. While it is natural for both men and women to pursue the best paying and highest status jobs, the Spanish labor market has a number of characteristics that make civil

[10] ANECA or "Agencia Nacional de Evaluación de la Calidad y Acreditación" is a Spanish agency created in 2002 to design/implement quality controls in higher education systems in Spain.

service jobs particularly attractive for women. The next few paragraphs will explain why.

Table 6 Percentage of Part-Time Jobholders in Spain

	2005	2006	2007
Female	24.2%	23.2%	22.8%
Male	4.5%	4.3%	4.1%

Source: European Economic Statistics 2008[11]

Spain has the fourth highest female unemployment rates of the EU-27 (10.9% for females, 6.4% for males); only in Croatia, Slovakia, and Greece are female unemployment rates higher than in Spain. The unemployment gap between men and women in Spain (i.e. 4.5%) is the second highest in the EU-27 (Greece is first with 7.6%). To put these figures in perspective, the average unemployment rate in the EU-27[12] is 7.8% for women and 6.6%, and in the US 4.5% and 4.7% for women and men respectively.

Second, women in Spain are considerably more likely than men to hold part-time contracts (see Table 6). This fact reflects that women face family obligations and are compelled to reduce their work hours even if that means lower earnings.

[11] The document can be downloaded at: http://epp.eurostat.ec.europa.eu/portal/page?_pageid=1073,46587259&_dad=portal&_schema=PORTAL&p_product_code=KI-30-08-410

[12] http://epp.eurostat.ec.europa.eu/portal/page?_pageid=1996,45323734&_dad=portal&_schema=PORTAL&screen=welcomeref&open=/&product=STRIND_EMPLOI&depth=2

In terms of salary, Table 7 shows there are important differences between women and men's earnings across all professional/educational categories.

Table 7 Average Income of Men and Women Across Occupational/Educational Categories (2005)

	Women	Men	Women	Men	Diff.
	Euro	Euro	USD	USD	
Management	41138	63968	64586	100430	56%
5-Year University Degree	26733	35930	41971	56409	35%
3-Year University Degree	23059	29843	36202	46854	29%
Vocational Studies	20837	29377	32714	46122	41%
Administrative Jobs	14701	20801	23081	32658	42%

Source: Encuesta de Estructura Salarial 2005 (INE).
http://www.ine.es/jaxi/menu.do?type=pcaxis&path=%2Ft22%2Fp133%2F2004-2005%2F&file=pcaxis&L=0&divi=&his

As Heilman and colleagues (2006) have argued, even after women succeed in attaining high status jobs, they are never free of the biasing effects of stereotypes (see Rudman & Glick 2001). As Heilman and colleagues put it: "… it is precisely at this point that processes arising from the prescriptive aspect of gender stereotypes are set in motion, with a different set of negative consequences for working women" (Heilman and Parks-Stamm 2007, p. 58).

From this description of the features of the Spanish labor market several ideas can be gathered. First, it is more difficult for Spanish women to be hired. Second, once hired they earn significantly less than their male counterparts. Third, making progress in the professional is more difficult

for women. My argument is that this labor market picture should encourage women to pursue and be successful in the public sector. Entry in the civil service is governed by a more explicit structure, it is presumably a merit-based process, and once a position has been attained workers do not need to fight for improvements as much as in the private sector. Salary decisions, vacation time, leave time, and the like are formally regulated; male and female civil servants benefit roughly the same. Finally, 5-year university graduates can compete for jobs within group A which offer an entry level salary of about 57.000 USD (€36.000) per year which is above the average income of workers with the same level of education (Table 7). In sum, these are good jobs in terms of pay, work conditions, and other social benefits. Women want them as much as men, if not more. By being employed in the public sector, women solve most issues that put them at a disadvantage in private employment. So, why is it that female applicants are failing exams that would lead to their recruitment?

Civil Service Exams

One purpose of this project is to offer empirical evidence illustrating the concrete or real-world manifestations of mechanisms discovered in controlled laboratory settings. There is ample experimental evidence confirming that women are evaluated less positively relative to men even when they provide the same evidence for competence. However, there is very little empirical evidence about what these mechanisms look like in the real world. Although experiments are ideal for establishing causal relationships, the next step should be to show how

these mechanisms operate in natural contexts. My approach was to select a natural context that is highly structured (and meets the scope conditions of the theory I use) and account for aspects that would be experimentally controlled in the laboratory – i.e. performance, qualifications etc. In the paragraphs below I will offer a summary description of the context of this study and the specific setting I selected. Second, I will discuss the advantages and limitations of the setting.

In Spain, a fixed number of government vacancies are announced annually for specific positions by level of qualification required. Access to any of these job openings requires passing several rounds of testing that typically involve face-to-face exams. After a brief probationary period, the highest scoring examinees become permanent government employees. Each round of exams lasts several weeks, takes place throughout the year, and requires the participation of various evaluating committees. Applicants are allocated to examination dates and evaluation committees according to a lottery based on a first random draw.

This study focuses on a specific group, namely Administradores Civiles del Estado[13]. ACE are senior officials with broad administrative knowledge and responsibility in areas such as budget, human resources, and contracts in the Spanish Administration. ACE official are in charge of drafting top-level government proposals, as well handling public policy. They hold management-level positions in different government offices and European Union agencies. The highest ranking ACE officials work closely with politicians and many become prestigious and influential politicians themselves (Crespo 2004).

[13] ISCO equivalent: Senior government official (four digit code 1120)

The ACE selection process is highly competitive. Prospective applicants prepare for these exams full-time for no less than 2-3 years and up to 5 or 6. Nonetheless, hard work does not always pay off; some applicants abandon the pursuit entirely after repeated failure. Sometimes applicants prepare for these exams by themselves but more often they study under the mentoring and supervision of a personal trainer. Trainers are ACE officials themselves and their job is to guide and coach prospective candidates as they prepare for exams and go through the recruitment process.

The ACE recruitment process was selected based on several considerations. First, data from exploratory interviews suggested that ACE recruitment is perceived to be unbiased, which is not the case with other public recruitment processes in Spain. In-depth interviews were conducted with 12 applicants (58% women) in May 2005 and multiple informal interviews were carried out between September and December 2005. Interviewed applicants shared the notion that ACE recruitment is merit-based and often compared it to other selection processes which, in their minds, are not based entirely on merit. For example, recruitment for the Diplomatic Corps is perceived to be highly biased in favor of male candidates. Requirements for these positions are framed as incompatible with gender assumptions around family and motherhood. As a result women are often (1) discouraged from entering the actual competition (see Table 3), and/or (2) confronted with awkward questions during the actual exams – i.e. women are indirectly asked about their life plans and family prospects. Judges construe job demands (e.g. traveling. living abroad etc.) as incompatible with assumptions based on stereotyped

images of women, their goals, and aspirations (i.e. women want children, stability etc.). As Heilman (1983) has argued, stereotypes about women create problems because there is a perceived lack of fit between women's presumed capabilities and the requirements for a given job (Heilman 1983). This is particularly true when women attempt to gain access to upper-level jobs, as the qualities believed to be necessary for these jobs are usually perceived as male-typed (Lyness 2002).

By contrast, the ACE selection process is perceived as being more gender neutral. Gender status theories do not presume evaluators are explicitly sexist or prejudiced; rather, they argue that under certain conditions gender implicitly affects the perception of competence in a way that usually (but not always) disadvantages women.

A second advantage of this context is that, although it is not a controlled environment in the strictest sense, the setting is sufficiently structured with the important advantage of real world consequences. Applicants do not generally differ in their ethnicity, nationality, or level of education. Since gender is the characteristic of interest here, variation in any of the above would complicate the interpretation of results.

Similarly, all exams take place under uniform conditions (physical location, format of the exam, duration etc.). which allows for repeated observation of the same event with gender (of applicants and judges) being the one aspect that varies. Importantly, evaluators know very little about applicants (e.g. name. date of birth etc.) precisely to facilitate objectivity - SCT and DST predict that gender will have more of an effect when no other status information is present in the setting; thus, evaluators

will be more likely to rely on status beliefs and stereotypes to orient their attitudes and behavior.

Fourth, the selection of applicants is not done by an individual evaluator but by committees of 5 judges. About 20 judges participate in SCA recruitment every year; in addition, judges vary from one year to the next. If decisions were made by a single evaluator, it would be hard to rule out the possibility that the outcomes observed are simply the reflection of one person's preferences.

Judges come from different backgrounds of the Spanish administration; between 50-60% are ACE officials themselves, and the rest are university professors, diplomats, judges etc. Committees are generally gender balanced; in terms of age, committees are composed of an even number of junior and senior members. These features tie in with the very tenets of this selection process, namely impartiality.

There have been a tremendous effort to make civil service recruitment merit-based. The demographic composition of committees is important because traditionally evaluating boards were largely male. In the last ten years, major efforts have been made to make committees more diverse and balanced specifically with regards to gender. The general idea underlying this move is that more balanced boards will arrive at more unbiased recruitment decisions.

Finally, this study examines processes at various rounds of testing; the range of ability among applicants is muted after round one since the lowest performing applicants have presumably been eliminated from the competition. In sum, although the setting I selected has the limitations described, it is not very far distant from the kinds of conditions one would

want to recreate in a controlled laboratory experiment, with the important advantage of having real world consequences.

CHAPTER FIVE
PART I: HYPOTHESES & METHODS

As outlined in the introductory chapter, this project has two distinct analytical sections. This chapter and the next will discuss hypotheses and methods (Chapter 5) and results (Chapter 6) for part one. In the following paragraphs I will offer a detailed description of the ACE recruitment process. Interview and observational data were crucial to gain an in-depth understanding of the hiring procedures governing ACE exams and formulate some of the study's hypotheses.

In this competition applicants go through a total of four qualifying exams that take place throughout the year in Madrid – i.e. May, June, September, and November. All exam sessions are public; observers are asked to leave the exam premises when judges deliberate and assign a score, which is made public shortly after – i.e. test results are posted on a bulletin board outside the exam room. Examining boards are composed of 5 judges whose votes are independent and have equal value regardless of their rank or seniority.

Exams 1, 2, and 4 are written exams although the exercises are not read by individual judges. Rather, applicants themselves read their exams out loud in front of a board of judges (henceforth "reading session") on a specific day and time assigned to them at random. Exam 3 is purely oral; applicants are given one hour to verbally rehearse four questions drawn at random from an official study guide composed of over one hundred topics. Judges have 15 minutes to ask applicants exam-related questions – this part of Exam 3 will be referred to as Q&A portion or simply Q&A.

In exam 1 applicants have 4 hours to write a general knowledge essay that they later read to the evaluating committee. Although essay questions are broad, applicants are specifically required to relate their answers to the appropriate topics of the official study guide. Exam 2 is a foreign language test where candidates are evaluated on their translation and listening comprehension skills. In exam 4 applicants are given four hours to discuss in writing the solution to several applied questions. Applicants may consult their books and materials when writing exam 4. Exams 1, 2, and 4 are similar that they involve a written part and a 20-30 minute public reading session, which is usually scheduled days or weeks after the written part of the exams. In exams 2 and 4, judges have 15 minutes to engage in a dialogue with applicants and ask them exam-related questions - exam 1 does not include a Q&A portion. Finally, exam 3 is purely oral; applicants are given one hour to answer 4 questions drawn at random from topics in the study guide. Candidates have 20 minutes to write out the outline that later guides their one-hour uninterrupted performance. Evaluators have 15 minutes to ask exam-related questions at the end.

Although this is what official procedures establish, the actual exams either (a) differed from the official version and/or (b) are characterized by additional features identifiable only by direct observation. I conducted direct observation of exam sessions in May 2005 and from September to December 2005. Two main ideas were gathered from this fieldwork. First, one might think that since exam 3 is purely oral this feature sets it apart from all others. Direct observation made it clear that this was hardly the case. Due to the highly memory-oriented nature of

exam 3 and the strict time constraints (i.e. applicants must devote exactly 15 minutes to each of the four questions), applicants' presentations end up being highly scripted, fast-paced, and matter-of-fact; applicants' verbal style is not much more natural than it is in the reading sessions of exams 1, 2, and 4. Second, direct observation permitted observing that the most important differences between exams are (a) the presence/absence of a Q&A portion and (b) the nature of the questions asked in the Q&A. These differences and their implications will be discussed thoroughly next.

About 70% of applicants were not asked any questions at the end of exam 4. As for the remaining 30%, the Q&A rarely lasted more than 5 minutes on average (SD = 2.7 minutes) and seldom involved the participation of more than one judge. In January 2006 I interviewed a female judge and a male judge who participated in the 2005 ACE recruitment process. In-depth interviews lasted about 90 minutes and consisted of a series of open-ended questions about the recruitment process. My interviews included questions such as evaluators' perception of exams (i.e. which exams seem more/less difficult to evaluate and why). how decisions are made when judges deliberate, and about the inconsistencies between official exam procedures and those that were actually followed. Both judges I interviewed confirmed that evaluators tacitly agree not to ask questions at the end of exam 4; apparently the purpose of questions is to measure applicants' spontaneity. Applicants typically consult what they wrote at exam 4 with their personal trainers and prepare in advance for potential questions that may come up during the reading session. Thus, questions asked at exam 4 do not necessarily measure applicants' problem-solving abilities, but rather the extent to

which applicants prepared in advance for potential questions. Interview data clearly suggests that the Q&A portion is viewed as an opportunity for judges to gauge applicants' unplanned reactions to questions. When the ability to measure this is undermined, questions are no longer important and thereby avoided by judges.

All applicants were asked questions in exam 2. Since exam 2 is a foreign language exam, questions were intended to provide opportunities for applicants to express themselves in a language other than Spanish. Questions in exam 2 were drawn from a list that evaluators had previously thought out; thus, questions repeated frequently and were somewhat scripted (i.e. "what is your favorite pet?"). These questions were generally posed by one of the evaluators, a language expert, brought specifically for assisting the committee in exam 2.

Finally, the Q&A portion of exam 3 was systematically used, involved all or most judges, and lasted an average of 12.7 minutes per applicant. The Q&A part of exam 3 was the one and only section of the entire selection process that involved an actual dialogue between judges and applicants.

Gender Typing of Exams & Degree of Interaction

Recall that status characteristics theory predicts that when gender is salient in task oriented settings, men will be evaluated more positively than women so long as the skills being assessed are perceived to be masculine or neutral. There are reasons to believe that in Spain, women are typically thought to be better at verbal skills in general and foreign languages in particular. For example, about 75% of university students

majoring in English, French, Spanish, and Italian philology are women. Similarly, more than 80% of college students majoring in translation and interpretation are women. While it may seem that these figures simply show that Spanish women have a strong preference for these disciplines; it also suggests that generalized beliefs exist in Spain regarding women's superior ability in foreign languages. Following Correll (2001) I assume that individuals need to feel competent about something in order to pursue it: "…while many factors certainly influence individual career relevant decisions and preferences, as a minimum, one must feel competent at the skills or tasks necessary for a given career in order to commit oneself to pursuing that career." (Correll 2001, p. 1700). In other words, women chose these college majors, at least partially, because they believe they are good at foreign languages. Correll (2001) demonstrated that self-assessments or "personal conceptions of task competence" (Correll 2001) are shaped by broadly shared beliefs that exist in the culture and affect individuals' self evaluations, which in turn shape future career orientations. The key point is that women's choices reflect more than their preferences; these choices suggest that in Spain individuals share the idea that women are better than men at foreign languages. The prediction then is that female applicants will not disadvantaged at all exams:

Hypothesis 1: the group advantaged at each exam will vary as a function of the gender typing of the task to be evaluated.

Hypothesis 1a: male applicants will score significantly higher than female applicants when the skills evaluated are perceived as masculine or neutral (i.e. exams 1, 3, & 4).

Hypothesis 1b: female applicants will score somewhat higher than male applicants when the skills evaluated are perceived as feminine (i.e. exam 2).

Recall from chapter 3 that the shifting standards model would predict in my civil service context, that female applicants will be judged according to more lenient standards the first round of testing but not in others. Although exam 1 serves as a screening or filter, this does not necessarily imply that this exam it is less important or less demanding than the others. In fact, exam 1 ranks second in importance after exam 3. These exams are not incremental in difficulty, rather, each round of testing evaluates different abilities, all of which are required to succeed at the job. In fact, exam 1 sets the ACE selection process apart from other civil service recruitment processes. Typically exams are extremely memory oriented. The ACE corps prides itself on being different and thereby needs a different type of applicant. Exam 1 is there to ensure that applicants cannot only memorize but can also write well and connect ideas in coherent and compelling ways. Although it comes first, Exam 1 is not the equivalent of the preliminary interview in other hiring contexts. Thus, my prediction is that female applicants will be evaluated according to a stricter standard this round of testing as well.

As I argued in the last section of Chapter 3, a setting needs not involve direct interaction for SCT and DST predictions to apply. Based on Ridgeway's work (1997) on interaction and Heilman's (2001) ambiguity thesis, I propose that contexts involving less structured interactions encourage actors to enact their gendered beliefs to a greater extent than settings were interaction is more rigid or minimal. Thus, while gender

status theory predictions will work regardless of the level of structure in judge-applicant interactions, the magnitude of gender differences in the outcome of interest (i.e. exam score) will be greater when actors interact more naturally and freely than when relate to one another in more scripted ways. In less structured interactions, allowed and prohibited behaviors are not clearly defined. Thus, it is up to the individual to fill this gap or absence of rules, and actors will exercise such freedom in a gender biased manner. Thus, if evaluators have implicit preferences for male applicants, a less structured setting will enable them to behave in ways consistent with these beliefs more so than settings were interaction is more guided or structured.

Recall the thorough description of exams provided earlier in this chapter. Table 8 below summarizes some of the key ideas that need to be kept in mind to understand the motivation of subsequent empirical predictions.

Table 8 Degree of Interaction Characterizing
Each Section of all Four Testing Rounds

	Exam	**D.I.S.**	**Q&A**	**D.I.S.**	**Total**
Round 1	Read	+	NO	-	+
Round 2	Read	+	YES	+	++
Round 3	Oral	+	YES	++	+++
Round 4	Read	+	NO	-	+

Note: D.I.S. = Degree of Interaction Score

In assigning each testing round a degree of structure of interaction score I have considered (a) characteristics of the actual exam (i.e. “Exam”

column in Table 8) and (b) features of the Q&A part (i.e. "Q&A" column in Table 8). Even though exams 1, 2, and 4 are read and exam 3 is oral, they all receive only one plus sign because, as argued earlier, applicants' speech style is similarly scripted across all four exams. Only exams 2 and 3 have a Q&A portion at the end. The Q&A of exam 2 is more artificial than that of exam 3; thus, exam 3 receives two plus signs. According to this classification, exam 3 is the most interactive, followed by exam 2, and exams 1 and 4. The essential point is that exam 3 is different from all others because applicants and judges engage in a fairly natural conversation, which is not the case in any other exam.

It has been argued that the degree of structure in applicant-judge interactions will impact the magnitude of H1a and H1b predictions. If so:

Hypothesis 2: the magnitude of gender differences in passing rates will be affected by the degree of structure characterizing a given exam.

Hypothesis 2a: small differences will be detected between male and female's passing rates in exams characterized by minimal applicant-evaluator interaction (i.e. exams 1 and 4).

Hypothesis 2b: larger differences will be detected between male and female's passing rates in exams involving a greater degree of applicant-evaluator interaction (i.e. exams 2 and 3).

Table 9 summarizes the above classification as well as all empirical predictions. According to status characteristics theory female applicants will be disadvantaged in all but exam 2 (i.e. exam 2 evaluates female-typed skills). By introducing the concept of degree of structure in interaction, I make a more nuanced prediction; namely, that the magnitude

of SCT predictions will vary as a function of structural features of the setting; small gender differences will be observed in exams 1 and 4 and larger gender differences in passing rates will be observed in exam 3.

Table 9 Exam Classification and Empirical Predictions

Exam	Exam Characteristics		Predictions	
	Gender Typing of Task	Degree of Interaction	Advantaged Group	Magnitude of Advantage
1	Neutral	+	Men	Small
2	Feminine	++	Women	Moderate
3	Neutral	+++	Men	Large
4	Neutral	+	Men	Small

Data & Methods

I gathered and examined data for 1476 ACE applicants (514 men and 955 women) who participated in ACE competitions between 2003 and 2005. Since 2003 the results of major civil service competitions are posted in PDF documents at official government websites. These data could be gathered and prepared for statistical analysis through an automated program. Publicly available information includes the applicants' names, their personal identification numbers[14], and their exam results at each testing round. In addition, these records contain the date and order in which candidates take exams.

[14] The Spanish equivalent of the Social Security Number, a unique identifier issued by the government to each Spanish citizen.

From this publicly available information I created a set of variables indicating identity (name, SSN), gender, exam scores, and the order in which applicants are tested. Applicants who fail an exam do not receive a numeric score; these applicants' measures were coded as missing. Finally, there are variables relating to the relative order in which applicants go through testing rounds (order is always assigned based on a first random draw), which have no theoretical importance and will therefore be excluded from the analysis.

The resulting data file consisted of multilevel data with three levels: candidate, exam, and year. To evaluate the hypotheses above, I used logistic regression clustered by applicant ID to control for the fact some candidates reenter the competition after having failed in previous years and, as such, observations are not independent.

A shortcoming of these data is that the available demographic information on applicants and judges is limited. Importantly this setting is highly controlled; there is little variation among applicants in terms of age, education, ethnicity, and other status information. In order to compete for these positions applicants must have a Bachelor in Arts or Bachelor in Science degree – the majority of applicants (about 75%) have a degree in Law, which is an undergraduate major in Spain. Thus, the initial pool of applicants is fairly homogeneous except for the gender characteristic of interest.

In the next chapter I will evaluate if SCT can account for the outcomes observed in this natural context. I have argued that women will be disadvantaged (i.e. score lower than their male competitors) in exams assessing male/neutral skills. In addition, I have proposed that the degree

of structure in applicant-judge interactions will impact the size of this disadvantage. In the next chapter I evaluate these hypotheses and discuss their implications.

CHAPTER SIX
PART I: RESULTS & DISCUSSION

Throughout this project I have argued that Spanish women are interested in attaining high status government jobs (i.e. 64% of the initial applicant pool are women) but that they fail exams that would lead to their recruitment at greater rates than their male competitors. There are various basic explanations for this outcome. Either female applicants are less skilled than male applicants, evaluations are biased in favor of males, or a combination of both.

Although it cannot be ruled out yet that female applicants might be under-qualified, this work examines processes at various rounds of testing; thus, it can be argued that the range of ability among applicants is muted after round one since the lowest performing applicants have presumably been eliminated from the competition. In addition, the current literature supports that evaluations are often biased in favor of men, except when the setting is female-typed.

In this section I will evaluate the adequacy of status characteristics theory to explain how male and female applicants perform in all four rounds of testing. I will also use what I call degree of structure in applicant-judge interactions to understand variations in the size of SCT predictions.

The following paragraphs will examine the following questions: (i) in what exams are female applicants disadvantaged, and (ii) what is the magnitude of such disadvantage.

Hypothesis 1: Sex-Typing of Exams

Recall from chapter five that hypothesis 1 is concerned with how female applicants fare at rounds of testing involving the evaluation of male/neutral and female-typed skills. SCT does not predict that women will always be evaluated less positively relative to men. Rather, SCT argues that men will be perceived as diffusely more competent than equally qualified women so long as the task evaluated is not one where women are stereotypically thought to excel (e.g. foreign languages). As argued in chapter five, the high presence of women in foreign language majors at Spanish universities suggests that generalized beliefs exist in Spain that associate women with excellence in foreign languages. Exam 2 of the ACE selection process involves assessing English and French skills. Thus, the prediction is that female applicants will not be disadvantaged in this particular exam. The rationale would be that evaluators are aware of broadly shared beliefs regarding women's superiority in foreign languages. Thus, judges will unconsciously expect female applicants to do better than males in exam 2 – i.e. judges' performance expectations for women will be high in the context of exam 2. Thus the prediction is that male applicants will pass at greater rates than female applicants across all exams except exam 2. The following model will permit evaluating such hypothesis:

$$y = a + b\,x_1 + c\,x_2 + d\,x_1\,x_2 + e$$

where y is the dependent variable pass, x_1 is applicant's sex (1=female), x_2 is exam 2. I predict that the main effect of sex (female=1) on the log odds

of passing an exam will not be significant but that the interaction effect of sex and exam 2 will positively impact the log odds of passing (see Model 1 in Table 10).

Table 10 Coefficients from a Logit Regression of the Log Odds of Passing an Exam on Gender, Exam 2, & Gender*Exam 2

	Model 1
Female	-0.303
	1.71
Exam 2	0.690
	(2.68)*
Exam2*Fema	1.029
	(3.07)*
Constant	0.016
	0.11
Observations	1372

Note: N = 1372. Observations are clustered by ID (N=424). Absolute values of Z statistics in parentheses.

The results in Table 10 (Model 1) suggest that female applicants have a higher probability of passing exam 2 but not others - although the negative main effect of gender on the log odds of passing an exam is not significant at conventional statistical levels. The interaction term shows that female applicants are about twice as likely to pass exam 2 than male applicants (1/exp[1.029-0.303]). As for exams 1, 3, and 4, female candidates are about 1.35 less likely to pass than their male competitors (1/exp[-0.303]). These results confirm that women do better than men at exams where female-typed skills are assessed, and somewhat worse than

male applicants when the abilities evaluated are masculine or neutral. But, are female applicants equally disadvantaged at exams 1, 3, and 4? Answering this question is the task I next turn to.

Hypothesis 2: Degree of Structure of Interactions

I have argued in previous chapters that the degree of structure of judge-applicant interactions varies across exams. If exams were ordered from greater degree of structure to lowest: exam 1, exam 4, exam 2, and exam 3. SCT predicts that women will score lower at exams 1, 3, and 4 because male/neutral-type abilities are assessed. I now turn to the question of whether or not women are equally disadvantaged across these three exams. If degree of structure in applicant-judge interactions matters, it should be possible to explore whether the negative impact of female on pass varies in magnitude across exams 1, 3, and 4. Based on my argument that less structured interactions will benefit high status groups and vice-versa, I predicted that the magnitude of female applicants' disadvantage will be greater in exam 3 than in exams 1 and 4.

Recall that exam 2 evaluates female-typed abilities, thus no gender penalty is expected for women in exam 2; rather, women will enjoy an advantage relative to men in this exam. Although I have argued that exam 2 is more interactive than 1 and 4, no predictions can be made as to the degree condition will impact the female advantage – i.e. this is the only exam where stereotypically female abilities are evaluated.

Table 11 Coefficients from a Logit Regression of the Log Odds of Passing an Exam on Gender, Exam, and Year

	Model 2	Model 3
Female	0.675	0.678
	(2.27)**	(2.21)**
Exam 1	-0.999	-0.868
	(3.82)*	(2.55)**
Exam 3	-0.412	-0.634
	(1.20)	(1.50)
Exam 4	0.519	0.672
	(1.13)	(1.10)
Exam1*Female	-0.729	-0.721
	(2.18)**	(2.11)**
Exam3*Female	-1.629	-1.693
	(3.87)*	(3.94)*
Exam4*Female	-0.959	-0.960
	(1.66)+	(1.63)+
Year 2003		-0.198
		(0.51)
Year 2005		0.251
		(0.74)
2003*Exam1		-0.064
		(0.15)
2003*Exam3		0.149
		(0.28)
2004*Exam4		-0.045
		(0.06)
2005*Exam1		-0.208
		(0.54)
2005*Exam3		0.453
		(0.97)
2005*Exam4		-0.454
		(0.68)
Constant	0.852	0.755
	(3.69)*	(2.53)**

Note: N = 1372. Observations are clustered by ID (N=424).

To evaluate this hypothesis I estimated the following model:

$$y = a + b\,x_1 + c\,x_2 + d\,x_3 + e\,x_4 + f\,x_1\,x_2 + g\,x_1\,x_3 + h\,x_1\,x_4 + e$$

where x_1 denotes applicants' sex (1=female) and x_2 x_3 x_4 are exams 1, 2, and 4 respectively. Because I argue that the negative effect of female is not invariant across exams the model also includes the interaction of sex and the exam dummies. As Table 11 (Model 2) shows, there is a positive main effect of sex on the log odds of passing. However, this main effect is adjusted downward for all exams except the baseline exam 2 by the value of the exam and gender interaction coefficients. Model 3 in Table 11 controls for year's effects and the interaction effect of years and exams. The interaction terms of gender and exam in Model 3 show that women and men have roughly equal log odds of passing exam 1 (1/[exp(-.72+.68)]), women are more likely than men to pass exam 2 ([exp(.68)]), and men are more likely than women to pass exams 3 and 4. The strongest gender effect is in exam 3, which men are roughly 2.76 times more likely to pass than women (1/[exp(-1.69+.68)]). As for exam 4, men are about 1.33 times more likely to pass than their female counterparts (1/[exp(-.96+.68)]).

Although statistically significant, differences between male and female applicants passing rates at exams 1 and 4 are small. In the case of exam 1 the difference is not substantively significant. Female applicants enjoy a considerable advantage in exam 2 which involves the assessment of female-typed abilities. The gender coefficient on the log odds of passing exam 2 indicates that women are just under twice as likely to pass

exam 2 as men. This is a bit inconsistent with what SCT would predict which is that women would do somewhat better (but not far better) than men in exam 2.

These results could raise the concern that since a high number of female applicants pass exam 2, selection effects may boost the magnitude of the negative effect of gender in exam 3. In other words, it could be the case that female applicants who are not sufficiently competent make it to round three due to the higher performance expectations evaluators hold them to in round two. This would be a fair concern if exam 2 was the first round of testing. However, note that applicants have already pass through a first filter, exam 1. Presumably the lowest performing candidates are eliminated at the first round. Exam 1 is highly demanding and requires that applicants have carefully studied the official study guide topics. Recall that exam 1 consists of writing an essay about a substantive topic and, importantly, connecting that topic to the appropriate subjects of the study guide. Thus my argument is that exam 1 is an effective first filter in this selection process. In addition, judges are highly compelled to be strict and demand high standards from applicants in exam 1. The duration of the selection process depends largely on the decisions made by evaluators in exam 1. Understandably, judges are rationally oriented to making the selection process as short and efficient as possible. Thus, they are motivated to select only those applicants that are perceived to have a real shot at winning the competition. This is true to all rounds but specifically so to round one where the applicant pool is larger.

In sum, the results above confirm that (1) as being female does not always impact pass outcomes negatively; rather, the advantaged or

disadvantaged group will vary as a function of the gender typing of the abilities evaluated like SCT predicts, and (2) classifying exams by the degree of structure of applicant-judge interactions is useful to explain variation in the magnitude of SCT predictions.

In the next chapters I will examine the processes at play in exam 3. I have argued that female applicants do much worse in this exam than in any other because applicants and judges interact more freely or to a greater extent in this round of testing. Next I will elaborate on this idea by providing a detailed account of the processes and behaviors at play in the Q&A part of exam 3. I have argued that it is the characteristics of this Q&A portion that makes exam 3 different from all others. I identified and quantified a behaviors that only occurred in the third round of testing making exam 3 Q&A less structured. This is negative for women because evaluators will have more freedom to act on their gendered attitudes and exhibit subtle discriminatory behaviors. In chapter seven I will specify a set of hypotheses that will help us understand how judges' behaviors reflect biases against and disadvantage female applicants.

CHAPTER SEVEN
PART II: HYPOTHESES & METHODS

This section focuses on the processes at play in the Q&A portion of exam 3 where judges take turns asking questions to each applicant for about 15 minutes. I have proposed that since interactions are less structured in this part of exam 3, judges will be more likely (in the absence of explicit rules) to use their own gender biased criteria to orient their attitudes and behavior. Biases in evaluations are typically studied in experimental settings where performance is experimentally controlled and gender is experimentally manipulated. Thus, observed differences in outcomes (e.g. being recommended, hired, or promoted) reveals differences in assessments due to factors outside objective skill, competence, or qualifications – in this case, the factor would be gender.

In natural settings perceptions of performances are revealed and become known much more subtly. For example, an employee reporting to his boss may not receive an explicit assessment of his/her input. Nonetheless, the boss will certainly form an impression of the employee's contribution in the course of their encounter. How does one gather the boss' assessment? First, an estimate could be obtained by self-report – i.e. asking the boss directly. Another option would be to focus on the content and style of the employer-employee interaction - i.e. verbal and non-verbal cues. The former is problematic and unreliable because implicit opinions might not be apparent to the actor. Also, social desirability issues pressure individuals to provide socially acceptable responses even if their true opinions differ.

If one were to instead focus on verbal and non-verbal cues, one could measure behaviors such as looking at/away from the speaker, fidgeting, nodding, cooperative and non-cooperative utterances and so forth. For instance, if the listener looks away while his/her partner is speaking it might indicate lack of interest. This in turn, could frustrate the speaker who may then become distracted and loose track of what's in his/her mind.

As the example above suggests, in natural contexts individuals are not usually confronted with immediate and tangible assessments, but rather with behaviors revealing positive or negative evaluations from others. Expectation states theory argues that inequalities in task oriented groups are due to the differential performance expectations members hold for themselves and one another. When group members hold high expectations for an actor they behave as though the actor's performance is likely to be useful to the group. Low expectations reflect the reverse. A performance expectation is a theoretical construct that is not directly observable. For these reasons, in this work I focused and measured behaviors that presumably reveal (i) the judges' performance expectations for applicants and (ii) the applicants' performance style. In other words, the interest is on behaviors that give away or signal something about evaluators' thought processes as they assess applicants, as well as applicants' displays of confidence or lack of confidence. Although my coding criteria will be thoroughly explained in the data and methods sections, I will offer a succinct review of some of the existing research on gender and non-verbal behavior and will then specify concrete empirical predictions.

Who Interrupts Whom?

Participants in conversations are expected to adhere to the turn taking system; fairness in conversations entails that only one speaker talk at a time (Marche & Peterson 1993). Interruptions are violations of the unwritten conversational contract and have been defined as instances of simultaneous speech that involve an intrusion into the structure of a speaker's utterance (West & Zimmerman 1983). Interruptions display rudeness and a lack of respect for the speaker. they restrict the rights of speakers as well as allow interrupters to control the topic of conversation and exert control and dominance over their conversational partner (Marche & Peterson 1993). By interrupting one's partner, one is in effect saying that the partners presence or input is not equal to one's own and hence can be overlooked. Interruptions have been interpreted as a subtle reminder of the others lesser worth (Smith-Lovin & Brody 1989; LaFrance 1992).

Although the language literature is quite consistent in considering interruptions as negative sanctions. I argue that this specific setting calls for even stricter conversational rules insofar as a subset of the participants are being evaluated and thus are in a vulnerable position. These interviews should be opportunities for applicants to showcase their knowledge and skills. Interruptions interfere with this goal by limiting applicants' floor time, and putting applicants' confidence at ease thereby crippling their ability to express ideas effectively. Interruptions presumably occur when judges believe applicants are not doing so well; thus, measuring interruptions should give us an accurate read of how judges assess an on-going performance.

I have argued that in less structured interactions (i.e. exam 3 Q&A) judges and applicants will be more likely to enact behaviors reflecting preconceived beliefs about gender. But recall from the theory section that SCT and DST make specific predictions about how men and women behave in this setting. Generally, SCT and DST predict that evaluators will assume that male applicants will produce more valuable contributions than female applicants. Then, SCT and DST predict, women will receive fewer opportunities to participate, will be treated with less deference, and stricter standards will be used to evaluate their performance.

In this context, these general predictions can be assessed by focusing on judges' interrupting behavior, which I argued displays disregard for the speaker and limits a speaker's opportunities to make contributions.

Having lived in both the United States and Spain, it seems clearly that interrupting is less counter normative in the latter. While it is true that conversational rules are culture specific and that in Spain interrupting one and other in informal conversations may be more socially acceptable, this setting is highly formal and as such strict conversational rules should also apply. In other words, setting specific rules should prevail over culture-specific norms; interrupting in this context is likely a norm violation. In fact, applicants are not interrupted most of the time (i.e. 73% n=938). In addition, while it might be the case that Spaniards are more prone to interrupting relative to individuals from other nations, the central point is whether male and female applicants receive differential interruptions. In other words, even if on average Spanish judges interrupt more frequently than hypothetical non-Spanish evaluators (i.e. mean levels may vary

across cultures), it is gender differences that matter here (i.e. the gender gap in interruptions could be similar across cultures). Finally, direct observation permitted detecting signs of stress in applicants when they were interrupted repeatedly. These responses were not measured systematically but included sighs, nervous laugher and the like, which confirms that applicants take interruptions negatively and not as the cultural norm.

Hypothesis 3: female applicants will receive more interruptions than male applicants.

The literature on gender an interruptions suggests that women are less likely to interrupt than men (Eakins & Eakins 1978; Smith-Lovin & Brody 1989). The following hypothesis evaluates this claim:

Hypothesis 4: male judges will interrupt applicants more often than female judges.

When do Interruptions Occur?

One of the main interests of this work is to identify the circumstances under which interruptions occur. By discerning when, the why can also be inferred, thus we would be in a position to gather judges cognitive processes or performance expectations which, as argued, are not directly observable. Interruptions presumably occur, at least partially, in response to applicants' behaviors that reveal the quality or style of their performances. In particular, I measured (i) applicants' pauses and (ii) applicants' utterance duration.

As argued in the theory chapter, double standards theory suggests that similar behavior will be interpreted differently as a function of salient status characteristics such as gender. When a behavior is displayed by individuals for whom there exist different expectations, its meaning may be interpreted very differently. Empirical evidence demonstrates that when actors differ in gender, the implications drawn from their behavior is quite different (Kunda, Sinclair, & Griffin 1997). For example the same demeanor may be characterized as laid-back when exhibited by a man, and as timid when exhibited by a woman. Thus, behavioral information is quite malleable, with its meaning varying depending upon what is expected. The nice feature of pauses and utterance duration is that, unlike interrupting behavior, both measures are susceptible to be interpreted in positive or negative ways. Applicants may pause because they don't know the answer to a question, because they are nervous, because they are thoughtful and so forth. It is up to the evaluator to make inferences about what causes applicants' to behave in such a manner. DST predicts that these inferences will be largely made on the basis of gender rather than, for instance, on objective performance. In other words, judges will be inclined to interrupt female applicants when they pause because pauses will be taken to project insecurity. Conversely, judges will not interrupt male applicants when they pause, as they will be seen as a sign of self-control and poise. If so:

Hypothesis 5: female applicants will be interrupted more often when they pause relative to male applicants.

Similarly, utterance duration might be viewed in a positive or negative light. Candidates could simply take more time to digress and hope that they come up with something relevant to say or, conversely more time could be used to provide a more detailed and accurate answer. Again, according to DST judges will associate the former with female applicants and the latter with male applicants. If so.

Hypothesis 6: female applicants will be interrupted more often when their answers last longer relative to the answers of male applicants.

Pauses & Utterance Duration

Even if hypotheses 4 and 5 are confirmed, it is entirely plausible that men and women pause and give lengthy answers for different reasons, in which case confirmation of the hypotheses above would not automatically imply that women are being treated unfairly.

As mentioned earlier, experts were hired to evaluate (a) questions' difficulty and (b) answers' quality. Having objective measures of these two will make it possible to ascertain whether exam judges are indeed using a double standard to evaluate applicants or, conversely, male and female applicants' similar behaviors have different causes. The following two hypotheses will be assessed:

Hypothesis 7: pausing will not hinder the objective quality of applicants' performance (thereby suggesting that pausing results from thoughtfulness not incompetence).

Hypothesis 8: answer duration will not hinder the objective quality of applicants' performance (thereby suggesting that speech duration results from knowledge not empty discourse).

If hypotheses 7 and 8 are confirmed, it will be possible to establish that a double standard to judge applicants is being used. In other words, if both male and female applicants pause and extend their speech because they are thoughtful and wish to showcase their knowledge carefully and at length, there is no reason why applicants should be penalized with interruptions when they exhibit these behaviors. If women are being interrupted more when their behavior could either signal carefulness or incompetence but we know for sure it does signal the former, it can be determined that it is judges' biases based on gendered assumptions that compelled them to make erroneous inferences leading to differential and unfavorable treatment of female applicants.

Number of Questions & Question Difficulty

DST argues that stricter standards will be used when evaluating female applicants or that more evidence of competence will be required from women to consider them as capable as men. In this context, this could mean that women receive more difficult questions from judges than male applicants. Since I gathered unbiased measures of question difficulty, it will be possible to evaluate the following hypotheses:

Hypothesis 9: female applicants will be asked more difficult questions than male applicants.

Hypothesis 10: female applicants will be asked more questions than their male counterparts.

Is Objective Quality Enough?

Gender status theories suggest that similar evidence of skill is interpreted differently for male and female applicants and that male applicants' performance will be evaluated more positively than that of their female competitors. Therefore,

Hypothesis 11: the objective quality of male applicant's answers will have a greater positive impact on the likelihood of passing the exam than that of female applicants.

Data & Methods

Evaluating the hypotheses above required using very different set of data that was obtained from covert observation[15] of live exam sessions. Many feel that covert observation is unethical. In this case, it was judged best for the wellbeing of individuals in the setting to proceed in this manner. These are public exams; as such, both applicants and judges are

[15] Summary of IRB Procedures: this research was governed by and passed US institutional review board but was not required to go through the Spanish equivalent. I obtained permission to conduct interviews and covert observation and taping of exam sessions. The justification for proceeding covertly was that not doing so would cause suspicion and stress among participants. Recording was necessary because note-taking proved insufficient and ineffective. Exam sessions were taped using an MP3 device unnoticed by participants. The recordings were daily downloaded and saved under password as voice files in my Macbook. Nobody except myself had access to the recordings. Transcriptions do not have attached names or any personal information which could give away the identity of the individuals whose voice was recorded. Recordings will under no circumstances be played in public or used in talks/presentations. All recordings will be destroyed well before the results of the study are published. Finally, I reviewed the regulations governing these exams and found nothing which indicating recording these public exams is illegal in Spain. Furthermore, I consulted these issues with two Spanish lawyers whose expert opinion is that the research design is in accordance with Spanish law. In sum, the proposed data collection methodology did not pose risks to participants nor is it illegal under Spanish law.

used to having an audience when exams take place. My presence in the setting was perceived as natural and routine. Second, telling participants about the purpose of me being there might have affected their actual behavior in the classic Hawthorne effect (Landsberger, 1958) thereby leading to lower quality data and, most importantly, affecting applicants' chances for employment. Similarly, briefing the judges on the research may affected their behavior.

Exam sessions were taped using a concealed MP3 device; recordings were then processed and transcribed. In addition, these data required extensive coding and cleaning before they could be analyzed. About 68% (n=83) of exam sessions were taped. Most were recorded between November and December of 2005, covering the full 8 weeks over which Exam 3 took place. The remainder exam sessions could not be taped due to various reasons. Sometimes the committee assistant would forget to invite observers inside the exam premises (n=8). There are no reasons to suspect that failures to remember this were intentional. Some candidates specifically requested that there would be no observers in the exam premises. In order to avoid interacting with future applicants before their exams, I stayed away from the exam room as much as possible. As a result, I failed to hear the committee's assistant invite observers. Third, I missed 17 exams for personal reasons. The latter two forms of missing data are likewise unproblematic, as there is no reason to think that I missed a nonrandom selection of exams. A more troubling source of missing data is the 11 examinees (8 females, 3 males) who explicitly requested that there be no audience in the room, a request that I of course honored. Nonetheless I checked that these 11 applicants were not any

more likely to fail or pass than the rest, which could have been a source of bias in my final dataset.

Of these 83 recordings, 4 had to be excluded from some parts of the analysis due to very poor sound quality – i.e. questions and answers could not be transcribed. A total of 33 and 50 recordings of male and female applicants respectively were used in this analysis. In the actual population of 126 applicants, about 32% were male. To obtain a proportional sample about 26 males (i.e. 40%) and 54 women should have been recorded. Since the interest of this project is to examine gender differences, male applicants were slightly over sampled so that the male group would not be too reduced (see Table 13).

Table 13 Sampling Details of Total Exam Sessions and Taped Exams Sessions, ACE Competition 2005

	Men	Women	Total
Total	40	86	126
Exams	32%	68%	100%
Taped	33	50	83
Exams	40%	60%	100%

Each of the 79 recordings contained (1) the actual one-hour rehearsal, and (2) a 15 minute Q&A portion where judges ask applicants questions related to part one. A total of 100 hours of exams were taped; of those, approximately 17 hours (Mean= 12', SD=2'9") of judge-applicant conversations were processed and used for this analysis.

Transcription & Coding

The Q&A segments contained (1) judge utterances, typically in the form of questions, and (2) applicant speech, usually in the form of answers. Other utterances included: clarifications. minimal responses. question tags etc. For clarity purposes I henceforth understand as questions all judges' utterances and as answers all applicants' utterances.

All audible questions in the 79 Q&A portions were transcribed - a total of 883 out of 1281 questions that could be counted, roughly 69%. In other words, some utterances could be identified as questions posed by judges but coherent transcription was not possible as the audio quality was too poor[16]. Transcription of answers was even more problematic for several reasons. First, applicants tend to speak with a lower tone of voice relative to judges; while for a given recording transcribing questions was feasible, transcription of the answers was not always possible. Second, questions tend to be shorter – even if a proportion of the words were inaudible, the question could still be inferred. Answers on the other hand were longer and a little less focused making it more difficult to make sense of them in the presence of inaudible segments.

A subset of recordings with the greatest sound quality was selected for transcription of the answers. Next I will describe and explain the steps I took to be sure my selection criteria did not introduce bias. All recordings were rated on a sound quality scale of 1 (very poor) to 3 (excellent). Those rated as 2.5 or above were selected for transcription – 45% of sample. Table 14 shows that selected recordings of male and

[16] The exam premises are located in downtown Madrid. Noise pollution is high and this and other unpredictable factors affected the quality of some recordings.

female applicants result in a reasonably balanced sample. In other words, men and women who fail and pass are represented in similar percentages in my sample of 35 applicants. These 35 applicants answered a total of 591 questions. Of these, 513 could be transcribed (about 85%).

Table 14 Sampling Details of Recordings Selected for Answer Transcription

		Males	Females
Pass	High Quality	11	13
	Total	24	28
	%	0.46	0.46
Fail	High Quality	4	8
	Total	9	19
	%	0.44	0.42

The coding process was three-fold with each phase requiring a tailored approach. A first category of items could be quantified in a straightforward manner (i.e. number of questions, speech time measured in seconds). The second group of items (i.e. quality and difficulty of answers and questions) required hiring qualified coders unaware of the study's hypotheses. These coders evaluated questions' difficulty and answers' quality on seven-point scales. Finally, a third type of event (i.e. interruptions, pauses) were coded following prescriptions drawn from a review of the existing literature and by establishing inter-rater reliability[17].

[17] Inter-rater reliability was established using 5 recordings chosen at random and comparing my results with those of my assistant. Each comparison was analyzed so as to arrive at very specific definitions and criteria for coding the event of interest. When 90% consensus was reached, I started coding all recordings from scratch.

Two experts (a female and male) were hired to evaluate questions on a seven-point difficulty scale (1 = "very difficult" and 7 = "not difficult at all"). Evaluators where selected based on their expertise in the subject matter of the exams – i.e. they are or exam trainers of prospective ACE applicants. I selected two people who had recently passed the selection process themselves (i.e. 2-3 years as ACE officials) because they would be more likely to be well acquainted with the study guide than judges who entered the ACE corps 15 or 20 years ago.

Coders were first given transcriptions of the questions. Answers were removed so that they would not impact the perception of the question (i.e. a hard question may appear easier if the applicant provided a brilliant answer). The gender of applicants and judges was unknown to expert evaluators who were also unaware of the study's hypotheses. Each of the two raters evaluated a total of 881 questions.

The two expert coders had different rating behaviors. While the male rater avoided extreme values of the seven-point scale; the female rater preferred these instead of values in between. However, the ratings of both male and female experts point to a similar pattern, namely the absence of differences in the average difficulty level of questions asked to male and female applicants.

According to the male rater, the average question difficulty for female applicants was 4.45 (SD = 0.98) and for male applicants 4.40 (SD = 0.97). As for the female rater, the average difficulty for female applicants was 5.02 (SD = 2.39) and for male applicants 4.82 (SD = 2.56). The direction of the ratings is similar; according to both, female applicants are asked slightly easier questions (1=very difficult, 7=not

difficult at all). Differences between male and female applicants are greater according to the female coder (i.e. 0.2 versus 0.06 for the male coder). But for neither coder were these small gender differences statistically significant.

After coding the questions for difficulty, hired coders were given a list of the questions and answers and were asked to evaluate the quality of answers on a seven-point scale (1=very poor quality and 7=very high quality). I arranged several meetings with both coders separately and made sure they understood the purpose of the job.

Quality refers to how well or how accurately does the answer address the question being asked. The male coder's average rating for female applicant quality was 4.66 (SD = 1.69) and for male applicants, 4.95 (SD=2.04). According to the female coder, the average answer quality for female applicants is 4.95 (SD=2.42) and for male applicants 5.52 (SD=1.98). Again, answer quality differences between male and female applicants are larger according to the female coder's ratings (i.e. 0.57 versus 0.29 according to the male coder).

However different, the ratings of both coders lean toward the same direction, namely male applicants provide, on average, somewhat better answers than female applicants. If the coders ratings had pointed to different directions it could have been problematic because averaging over the two would have cancelled out the effect. This was not the case in the study; thus, differences between the coders are not too problematic.

In the next paragraphs I will describe the dependent and independent measures obtained from the three-fold coding process outlined above.

Dependent & Independent Measures

Pauses & Speech Duration: Pauses are understood as brief interruptions of speech (2 or more seconds) at the beginning of the speaker's turn. Speech duration was measured in seconds and refers to the time an applicant takes answering a question.

Interruptions & Quasi-Interruptions: An interruption occurs when the talk of one person is intruded upon by the talk of another person. The mere presence of speech overlap does not itself constitute sufficient grounds for calling something an interruption. Interruptions may be regarded as such if the first speaker is unable to finish making a point or the topic is cut out short by the intrusion (LaFrance 1992).

A limit of some past research has been to ignore distinction among different types of interruptions. Most prior works have treated interruption as a unitary event. Nonetheless, researchers have suggested caution in assuming that the term interruption is well defined and non-problematic. Prior research has operationalized interruptions very differently: (1) undefined or broadly defined, (2) excluding back channels and minimal responses, (3) successful interruption (e.g. Kollock, Blumstein, & Schwartz 1985; Smith-Lovin & Brody 1989). In this work I tried to go beyond these and differentiated four types of speech overlap that can be confused with successful interruption. Table 15 below summarizes four types of situations that were often observed in my Q&A recordings. Only situation 1 constitutes an interruption as I have defined it in this work.

In situation 2, although the hypothetical applicant has not finished the judge intervention is presumably made with the intent to clarify the original question. Clarifications of the sort may sometimes be disrupting

for applicants; nonetheless, it seems plausible that judges interrupt to clarify in order to help applicants. Sometimes applicants request that something be clarified to them. Both volunteered and requested clarifications were coded separately as their meaning could be open to interpretation.

Table 15 Examples of Interruptions and Quasi-Interruptions

Situation 1: Interruption	Situation 2: Clarification
J: when was the communist party made legal? A: in 1977 when… (unfinished speech) J: but what specific day and month? A: April 9th	J: when was the communist party made legal? A: In… (<1/2" pause. unfinished speech) J: I mean the Spanish Communist Party. A: April 9th 1977
Situation 3: Cooperative Overlap	Situation 4: Follow Up Question
J: when was the communist party made legal? A: In 1977 J: Uh. uh (cooperative overlap) A: April 9th to be specific.	J: when was the communist party made legal? A: in 1978 (< 1" gap) J: what specific day and month? A: April 9th

Note: "J" stands for judge and "A" stands for applicant.

Active listening can lead to simultaneous talk without being interruptive (West & Zimmerman 1983). Situation 3 exemplifies this case. Minimal responses such as "uh, uh" "yes" "aha" were not considered interruptions.

Sometimes judges asked a question right after an applicant finished answering. How these affect candidates is hard to determine. On the one

hand, it may convey a bit of impatience from the judge. But also, a fast-paced Q&A could suggest that the evaluator is satisfied with the answers given and wants to move on to the next swiftly. These events I have called quasi-interruptions were coded separately.

Answers' Quality & Questions' Difficulty: finally, the two expert raters evaluated a total of 881 questions (1=very difficult, 7= not difficult at all) and 453 answers (1=very poor quality, 7=very good quality). I averaged over the two coders and used the resulting difficulty and quality measures for analysis.

Method & Descriptive Statistics

Regression models were used to assess the impact of applicants' gender, pauses, and speech durations on judges' interruptions. A second set of models evaluate the impact of applicants' pauses and speech utterances on unbiased measures of answer quality. Third, I examined whether gender affected the quantity and objective difficulty of judges' questions. Finally, a regression model was used to determine the effect of gender, and objective answer quality on the likelihood of passing the exam. In these models, standard errors were clustered by applicant ID to correct for non-independence.

Tables 16 through 19 provide some descriptive statistics for the variables used. The level of analysis is question-answer (except in the cases specified above). Judges made a total of 1282 questions (462 to male applicants and 819 to female applicants). Male judges asked the majority of the questions (about 78%). The rest, 284 (22%) were asked by female evaluators.

Male judges clearly dominate the Q&A; meaning, it is not always the case that there are a disproportionate number of male judges. In fact, evaluating committees tend to be balanced in terms of gender (i.e. about 63% of applicants had a gender balanced committee; 7% had a female-dominated committee; 30% had a male-dominated committee).

Table 16 Descriptive Statistics of Main Variables

		N	Mean	S.D.
Gender	Women	50	60%	
	Men	33	40%	
Questions	Women	819	64%	
	Men	462	36%	
Answer Time	Women	819	2.44	2.15
	Men	462	3.35	3.55
Difficulty	Women	535	4.73	1.47
	Men	346	4.61	1.55
Quality	Women	275	4.81	1.69
	Men	178	5.21	1.39

Male applicants were interrupted in about 20% of the questions (N=91) while women were interrupted in about 30% (N=252). About 88% (n=71) of interruptions directed at male applicants were single interruptions while 12% (n=20) were multiple – applicants are interrupted more than once while attempting to answer the same question. As for female applicants. 66% (n=165) were single interruptions while 34% (n=87) of them were multiple.

Table 17 Summary of Judge Interruptions
by Applicants' Gender

	Interruptions	Total
Women	31%	100%
	252	819
Men	20%	100%
	91	462

Table 18 Summary of Applicants' Pauses
by Applicant's Gender

	Pause	Total
Women	15%	100%
	126	819
Men	17%	100%
	78	462

Table 19 Summary of Applicants' Pass/Fail
In Exam 3 by Applicant's Gender

	Exam 3 Outcome		
	Fail	Pass	Total
Female	40%	60%	100%
	20	30	50
Male	27%	73%	100%
	9	24	33

CHAPTER EIGHT
PART II: RESULTS & DISCUSSION

In the first part of this project (i.e. Chapters 5 & 6) I proposed that women suffer greater discrimination in exam 3 than in others because this round of testing allows a less constrained interaction between evaluators and applicants. In Chapter 7, I offered a set of empirical predictions to determine whether this general claim has empirical support. Simply put, I asked: if the features of exam 3 Q&A are pushing judges to act more on their gendered expectations, what behaviors would reveal it?

Next I asked various specific questions (hypotheses 3 to 11); the answers to these predictions should help us understand the specific mechanisms that are putting female applicants at a disadvantage relative to male applicants only in exam 3. Recall from the theory chapter that double standards theory argues that a different and more strict criteria will be used to evaluate women in mixed-sex settings. My results suggest that faced with similar information, evaluators will be less forgiving with female applicants than with male applicants. Furthermore, I will demonstrate that behaviors that are interpreted as indicative of lack of ability in female applicants are in fact uncorrelated with objective skill.

Hypotheses 3 & 4: Who Interrupts Whom?

I argued that male judges would interrupt female applicants more than male applicants. As explained in previous chapters, the literature on gender, status, and interruptions says that high status actors will be

entitled to and exhibit more dominant behaviors. Interruptions are seen as displays of status. Thus, judges, particularly male judges, will be more likely to interrupt female applicants than male applicants. If so,

$$y = \alpha + \lambda x_1 + \varphi x_2 + \varpi x_1 x_2 + e$$

where y is the independent variable judge interruption. x_1 is a dummy variable for applicant's gender (1= female applicant); x_2 is a dummy for judge's gender (1= female judge), and their interaction term.

The results of Model 4 are summarized in Table 20 and show that female applicants are more likely to be interrupted than male applicants except when a female judge is interviewing them. The magnitude and sign of coefficient λ indicates that female applicants interviewed by male judges will be interrupted roughly once every two questions ($\alpha + \lambda = 0.56$ per question) while male applicants interviewed by male judges will be interrupted once every four questions ($\alpha = 0.26$ per question). Female applicants appear to enjoy a greater advantage than their male counterparts when female judges are the ones interviewing them ($\alpha + \lambda + \varphi + \varpi = 0.16$ versus $\alpha + \varphi = 0.11$ for males interviewed by females). These results confirm what other authors have previously found – i.e. men interrupt women most of the time.

Hypothesis 5: When Do Interruptions Occur? Pauses

My focus on interruptions is theory-driven, meaning, the purpose of this research is not to settle academic debates on the topic. Interruptions are penalties that judges administer to applicants presumably when

applicants are not performing as expected. Judges interrupt if applicants make mistakes, beat around the bush, do not sound convincing, take too long to answer etc. So interruptions provide valuable information about judges' unobservable performance expectations for applicants.

A question of interest is then: when do judges interrupt? when do judges react negatively to what applicants say? Hypotheses 4 and 5 will help establish when interruptions occur and why. Recall from chapter 4 that I measured several applicants' behaviors, particularly pauses and speech duration. The literature on nonverbal cues tells us that these behaviors are important but that their actual meaning is not univocal. Pauses may reveal poise or insecurity; speech duration could originate in knowledge or lack of fluency. If judges interrupt when confronted with such behaviors, it can be inferred that they thought applicants were insecure and under prepared.

I built two regression models (Models 5 and 6) which will permit assessing when male and female applicants are interrupted. Knowing this will help infer why judges interrupted (i.e. they thought the applicant was unskilled) or did not (i.e. they thought the applicant was calm). Model 5 examines the effect of applicants' gender and pausing behavior on the likelihood of being interrupted:

$$y = \alpha + \lambda x_1 + \varphi x_2 + \varpi x_1 x_2 + \sigma x_3 + \varepsilon x_1 x_3 + e$$

where y is the independent variable judges' interruptions. x_1 is a dummy variable for applicant's gender (1=female applicant); x_2 is a dummy for judge's gender (1= female judge), and their interaction term. The model

includes applicant pauses x_3, and a second interaction term of pause and female.

As Table 20 shows, the patterns found in Model 4 still hold – i.e. female applicants are interrupted twice as much as male applicants when interviewed by male judges (females, $\alpha + \lambda = 0.52$ versus males, $\alpha = 0.26$). The coefficient for the main effect of pause σ is not significant and has a positive sign thereby suggesting that male applicants are not penalized with interruptions when they pause. In contrast, female applicants are interrupted ($\varepsilon = 0.18$ more for every pause made) when exhibiting the same behavior.

These results indicate that judges probably believe that women pause because they are unsure while men pause because they are cautious or thoughtful. The real question then becomes if judges' inferences are based on the objective performance of an applicant or on the applicant's status attributes. In other words, do male and female applicants pause for different reasons as the interrupting behavior of judges suggest? Or conversely, do judges interpret similar behavior differently as a function of the applicant's gender? I will return to this question when discussing hypothesis 7.

Hypothesis 6: When Do Interruptions Occur? Speech Duration

Like pauses, applicants' answer length reveals something about applicants' performances. Interpreting pauses is not precise because they can either signal confidence or insecurity. Similarly, speech duration can be perceived in different ways. Applicants may keep the floor because they are knowledgeable or because they are unsure and simply wander

hoping to say something relevant. If judges believe applicants give a lengthy answer because they are learned, it is safe to assume that judges will refrain from interrupting the applicant. On the other hand, if judges think applicants are giving unnecessarily long answers and going off on a tangent, judges will interrupt them so as to not waste limited Q&A time. To evaluate how judges respond to the length of applicants' utterances, Model 6 incorporates two more terms relative to Model 5, namely applicants' speech duration and the interaction of gender and speech duration:

$$y = \alpha + \lambda x_1 + \varphi x_2 + \varpi x_1 x_2 + \sigma x_3 + \varepsilon x_1 x_3 + \gamma x_4 + \psi x_1 x_4 + e$$

where y is the independent variable judges' interruptions, x_1 is a dummy variable for applicant's gender (1= female applicant); x_2 is a dummy for judge's gender (1= female judge), and their interaction term. The model includes applicant pauses x_3, and a second interaction term of pause and female, as well as x_4 which refers to answer duration, and a third interaction term of answer duration and gender.

The coefficients in Table 20 (Model 6) suggest a similar pattern than that observed in Model 5. While male applicants do not get penalized for keeping the floor, female applicants are more likely to be interrupted the longer they keep on talking (ψ = .005 per second). Female applicants (not male) are still penalized for pausing (ε =.17). The main effect of gender is still positive and large although the term is not significant at conventional statistical levels. This suggests that although female applicants may be disadvantaged in general relative to males, women are

punished when they exhibit particular behaviors such as pausing or giving lengthy answers.

Table 20 Regression Coefficients and Standard Errors for a Model of Interruptions (clustered by applicant ID)

	Model 4	Model 5	Model 6
Female Applicant	0.294	0.258	0.129
	(4.44)*	(3.74)*	-1.56
Female Judge	-0.146	-0.146	-0.134
	-1.47	-1.47	-1.34
Female_A*Female_J	-0.242	-0.232	-0.259
	(2.05)**	(1.97)**	(2.19)**
Pause		0.004	0.01
		-0.05	-0.14
Pause*Female		0.182	0.17
		(1.99)**	(1.86)+
Answer_Time			-0.001
			-1.04
Answer_Time*Female			0.005
			(3.06)*
Constant	0.259	0.258	0.293
	(5.04)*	(4.79)*	(4.59)*
Observations	1281	1281	1281
Applicant ID	83	83	83

Absolute value of z statistics in parentheses

These results are interesting because scholars have tried determine whether status hierarchies in task groups are based on factors related to performance or on behavioral dominance. These results align well with Ridgeway's findings about the relationship between nonverbal behavior and status (see Ridgeway 1987). Ridgeway provided empirical evidence

that status is based primarily on expectations about task performance rather than on behavioral dominance. In my context, this means women are not being interrupted just because they are women and thus susceptible to be controlled or influenced. Rather, judges interrupt women when women's behavior can be interpreted to confirm judges' preconceived gendered schemas. Simply put, if judges believe women, on average, are less competent than men, ambiguous evidence such as pausing is more readily seen as confirmatory of their preconceived ideas about women's competence. As a result, judges interrupt women when they exhibit certain behaviors rather than in a random manner just to show dominance or control over them.

Hypotheses 7 & 8: Gender Differences or Gendered Interpretations?

The key question is whether judges' decisions to interrupt are justified or not. In my discussion above I assumed judges' decisions are biased based on the extensive literature on biases and evaluations. But technically I have not ruled out the possibility that similar behavior of women and men (i.e. pausing, lengthy answers) has a different origin. In other words, it is plausible that male applicants pause because they are calm and women pause because they are unskilled. Although the assumption of unbiased evaluations is inconsistent with the literature, the data collected for this project can offer empirical evidence to clarify this question. This is the task to which I next turn.

Recall that expert coders were hired to rate the quality of applicants' answers. Since I have objective measures of skill, it will be possible to disentangle the above questions empirically. Model 7 will help

establish whether applicant's gender, pauses, and answer duration have any effect on objective answer quality (controlling for question difficulty).

$$y = \alpha + \lambda x_1 + \varphi x_2 + \varpi x_1 x_2 + \sigma x_3 + \varepsilon x_4 + e$$

where y is the independent measure answer quality, x_1 is a dummy variable for applicant's gender (1=female applicant); x_2 refers to applicant number of pauses; third, the model includes an interaction term female and pause, x_3 is speech duration, and x_4 is question difficulty.

Table 21 Regression Coefficients and Standard Errors for a Model of Answer Quality (clustered by applicant ID)

	Model 7
Female Applicant	-0.334
	-1.34
Pause	-0.476
	(1.85)+
Pause*Female	0.282
	-0.88
Answer_Time	0.006
	(1.98)**
Question Difficulty	0.085
	-1.55
Constant	4.631
	(13.25)*
Observations	378
Applicant ID	35

Table 21 summarizes the results for this model of answer quality. Although female has a negative effect on quality, the coefficient is not

statistically significant. Pausing behavior affects the quality of male answers not that of female applicants' answers – the main effect is significant and has a negative sign ($\varphi = -0.48$). Interestingly, this suggests that when men pause it is probably due to lack of skill or nervousness. Thus, pauses in male applicants are negatively correlated with objective answer quality. Using more time to reply to a question positively impacts the objective quality of both male and female answers ($\sigma = .006$). This suggests that when applicants provide lengthier answers they do so because they have relevant things to discuss thereby producing a higher quality response. Taken together these results show that pausing should not be viewed as a sign of weakness in female applicants. Furthermore, it is for male applicants that pauses are negatively correlated with the objective quality of their responses. If anyone, it is male applicants that should receive penalties (i.e. interruptions) for pausing. These findings illustrate how a double standard can operate in a natural setting. Behavior such as pausing or speech duration can easily be interpreted in a positive or negative light. As these findings suggest, exam judges prefer to give male applicants the benefit of the doubt (and not interrupt them); while denying the same treatment to women even when, objectively, it should be the other way around.

Having objective measures of answer quality was essential to clarify whether real gender differences among applicants or judges' gender biased assumptions are responsible for judges' differential treatment of male and female applicants. Previous research has rarely offered this kind of detailed information. Thus, demand and supply side mechanisms could hardly be adjudicated. At the same time, experiments

have demonstrated that evaluations are biased but laboratory settings are artificial. This means that the mechanisms found may unfold very differently in complex settings. This analysis convincingly proves that men and women are being treated differently when they behave similarly and for similar reasons. I focused on judges' interruptions because it is a behavior that could be observed and operationalized, and because as I have argued throughout this discussion interruptions can be interpreted as nothing other than negative sanctions in this setting. This doesn't mean that differential treatment is limited to this specific behavior nor that interruptions are the most important type of penalty.

Interruptions are consequential in substantive ways (i.e. they stress applicants, they make them tired and so forth) but, in the larger picture, interruptions are merely an example of how evaluators react differently to identical behaviors exhibited by applicants who differ in their status characteristics. With this I would like to stress that is probably an array of judges' behaviors that were not (or could not be) measured that would indicate a similar pattern.

Hypotheses 9 & 10: Do Questions Differ for Men and Women?

In the paragraphs above I showed that judges penalize female applicants by interrupting them when they are trying to answer a question and that there is no apparent justification for these interruptions vis a vis answer quality. In this section I will evaluate other behaviors that lead to support for differential treatment of men and women by judges. I argued that because judges will use harsher standards to evaluate women, they will ask them more (hypothesis 10) and more difficult questions

(hypothesis 9) . Double standards theory posits that when women provide evidence of skill, evaluators tend to scrutinize that evidence because it is inconsistent with their prior expectations. In this context, this might mean that women receive a greater number of questions than male applicants in the Q&A portion. I assess this with a simple regression model of the following form:

$$y = \alpha + \lambda x_1 + e$$

where y is the independent measure total number of questions and x_1 is a dummy variable for applicant's gender (1= female applicant).

Table 22 shows that statistically significant gender differences were found in the number of questions male and female applicants were asked - women receive about 2.5 more questions than do males (baseline=14). This suggests that judges are more demanding when interviewing female applicants. This relationship stays even when a measure of answer quality is fit in the model (Table 22, Model 9). In other words, if better answers would lead to fewer questions, it could be argued that more questions are asked when judges believe answers are not satisfactory and thus further proof of ability might be required. These data show that this is not the case. Quality does not have any effect on the fact that female applicants receive a greater number of questions. Women who perform as well as men receive on average 4 more questions from judges than their male counterparts.

Table 22 Regression Coefficients and Standard Errors
for a Model of Total Number of Questions

	Model 8	Model 9
Female Applicant	2.50	4.36
	(1.42)+	(1.87)*
Answer Quality		1.10
		1.34
Constant	14.12	8.99
	(1.11)**	(7.04)**
Observations	83	83

Double standards theory would suggest that, since stricter standards will be used to judge women's competence, exam judges will ask female applicants harder questions thereby raising the standard relative to men. Model 10 evaluates the effect of applicant's and judge's genders on my objective measure of question difficulty.

$$y = \alpha + \lambda x_1 + \varphi x_2 + e$$

where y is the independent measure question difficulty (1=more difficult; 7=less difficult), x_1 is a dummy variable for applicant's gender (1= female applicant); x_2 is judge's gender (1=female judge).

Table 23 Regression Coefficients and Standard Errors for a Model of Question Difficulty

	Model 10
Female Applicant	0.117
	-1.00
Female Judge	0.309
	(1.93)+
Constant	4.586
	(50.22)*
Observations	881
Applicant ID	79

As can be seen in Table 23 this hypothesis was not supported by the data. Although the coefficient for female goes is in the predicted direction (i.e. positive sign indicates that women receive harder questions), its standard error is quite large, and consequently not significant according to conventional standards. Since applicant's gender is not significant it cannot be argued that female applicants receive more difficult questions than their male competitors. Female judges ask both male and female applicants harder questions.

A model was estimated that included an interaction term female applicant and female judge but it was not statistically significant. This evidence suggests that both male and female applicants receive similarly difficult questions from male judges and harder questions from female evaluators.

These results suggest that judges' double standards are reactive rather than proactive. In other words, judges may not actively raise the standard by adjusting the difficulty level of questions depending on the applicant's gender. Rather, judges react differently when confronted with very similar information provided by male and female applicants (i.e. pauses and speech duration). Understandably it requires more cognitive resources to adjust the difficulty of a question based the status attributes of those being evaluated.

SCT and DST argue that these are processes that occur out of awareness, thus it is reasonable to assume that behavioral responses are also spontaneous rather than calculated. Weighting and adjusting the difficulty level of questions requires effort, it is possible that while double standards apply to other more involuntary behaviors such as interrupting, it does not impact conscious behaviors such as tuning the difficulty level of questions as a function of applicant's gender.

Hypothesis 11: Is Male and Female Competence Perceived Similarly?

DST argues that the same evidence of competence and ability is perceived differently depending on whether such evidence comes from men or women. In this context this theory would lead us to anticipate that that even when female applicants provide answers as good as their male counterparts, their chances of passing the exam would be lower. The logistic regression model specified below evaluates the hypothesis that objective performance quality is weighted differently by evaluators based on applicant's gender:

$$y = \alpha + \lambda x_1 + \varphi x_2 + \varpi x_1 x_2 + \sigma x_3 + e$$

where y is the independent variable pass (1=pass; 0=fail). x_1 is a dummy variable for applicant's gender (1= female applicant); x_2 is answer quality (1= female judge), the interaction term of female and answer quality; and x_3 question difficulty.

Table 24 Coefficients from a Logistic Regression of the (Log odds) of Passing Exam 3 on Gender, Answer Quality, & Gender*Quality

	Model 11	Model 12
Female Applicant	-.23	1.04
	.26	.79
Answer Quality		.27
		(.13)*
Female_A*A_Quality		-.25
		(.15)+
Question Difficulty		-.18
		(.08)*
Constant	1.06	.43
	(.17)**	.76
Observations	453	378

Table 24 below summarizes two models, first a simple model to evaluate the impact of gender on the log odds of passing exam 3, and second, a the more complex model specified above (Model 12). The results of Model 11 indicate that women are disadvantaged in exam 3 although the coefficient is not significant. Model 12 shows that objective quality increases the odds of passing the exam for male applicants only.

The coefficient for women is preceded by a negative sign and it is significant at 10%. It is intuitive to think that the greater the quality of an applicant's responses the greater his/her chances of passing the exam. If this were the case for both male and female applicants, answer quality would have a positive effect on the chances of passing for both men and women examinees. This results provide evidence that this is not the case.

It is true that the Q&A does not represent the entirety of the exam, only a part of it. Nonetheless it is safe to assume that performances in the two portions (one hour rehearsal and Q&A) are correlated. In other words, applicants who do well in the Q&A have probably done well in the first part of the exam. Thus, quality in the Q&A should be a good proxy for candidate's performance. These results suggest that evaluators use answer quality as a reliable indicator of competence for male not female applicants.

The results discussed above suggest clear evidence of a double standard when evaluating male and female applicants. Hypothesis 7 was disconfirmed, which indicates that judges do not consciously make it more difficult for female applicants. Judges interrupt women more and ask them more questions but these are spontaneous behaviors. Formulating questions of varying degrees of difficulty demands a more conscious effort. Judges' double standards revealed by interruptions occur in reaction to available evidence for applicants (pausing, speech length). This is consistent with the gender and status literature which points out that status characteristics shape performance expectations subconsciously.

CHAPTER NINE
SUMMARY & CONCLUSIONS

Empirical evidence demonstrates that sex segregation in employment is still high and several explanations have been proposed to explain it. However, almost all empirical research on sex segregation is based on data of people who already have jobs. Thus, it has been difficult to identify mechanisms operating at the point of hiring. Laboratory experiments have made important contributions about what these mechanisms are but still we need to know what they look like outside artificial settings.

This project represents one of the first efforts to combine the best of both worlds: real data on employer practices that are sufficient to identify micro level mechanisms. The results of this research indicate that the causal mechanisms discovered in controlled environments can also be found in more complex contexts. Furthermore, this work illustrates the usefulness of using theories developed in laboratory settings to guide research in real-world situations.

Gender status theories have argued that since men are perceived as diffusely more competent and skilled than women, concrete men will also appear to do things better than equally qualified women when gender is salient. In this project I have described and examined a setting that permits evaluating these claims in a natural environment. The hiring process I analyzed involves exams where both neutral and feminine skills are assessed. I found that female applicants do better at exams involving verbal skills, which are stereotypically viewed as female, while men do

better all other exams, which involve the assessment of more neutral abilities.

The results presented here also suggest that the degree of structure in interaction moderates the effects of salient status characteristics. Researchers have argued that the sex-categorization that takes place in interaction prompts the use of gender stereotypes to guide attitudes and behavior. I have shown that the degree of structure in applicant-judge interactions will impact the extent to which actors in the setting will be allowed to act on their gendered assumptions. Specifically, I have proposed that settings where interaction is minimal or less structured will leave it to the individual's choice to exercise behaviors attuned to his/her gender beliefs. Contexts where interaction more structured will have the opposite effect; namely, individuals in the setting will have more limited opportunities to behave according to their gendered expectations. In this work I demonstrated that women score significantly lower than men only in exam 3, the only round where applicants and judges interact in a less structured manner. In sum, the degree of structure condition has helped make more accurate predictions about the magnitude of the advantage or disadvantage that men and women face in this setting.

The second part of this research relies on more detailed data on the interactions of applicants and judges in the Q&A part of exam 3. I have examined judges and applicants behavior that reveal aspects of their evaluations and performances respectively. Judges performance expectations are not directly observable. However, I explained that measuring interruptions should reveal the judges' cognitive processes as they assess job applicants. For example. if judges believe the responses of

applicants are more relevant and interesting, judges will refrain from interrupting them. Similarly, if the contributions of applicants are perceived as less valuable, judges will have less tolerance for long speeches and will interrupt more often. Thus, an analysis of judges interruptions should provide a good read of judges' reactions as they evaluate applicants. Interruptions are a highly useful measure because, as argued throughout the project, they cannot be interpreted positively in this context.

I found that women are interrupted more relative to male applicants and that these interruptions seem unjustified. Even though interruptions may not impact women's performance directly, it is likely that other judges believe interruptions are deserved. Recall from chapter 3 that reward distribution affects performance expectations. What this means in my setting is that other judges will see these interruptions as legitimate and will infer that female applicants are under qualified.

In this project I have demonstrated that female applicants are treated worse even when there is no basis for linking their behavior to objective lack of ability. Also, this research demonstrates that male applicants do not receive these negative sanctions and that a much more lenient standard is used with them – even though behaviors such as pausing are negatively correlated with objective quality answers in male applicants.

Having behavioral measures from both evaluators and applicants, as well as objective measures (i.e. answer quality and question difficulty) made it possible to identify the use of double standards in the treatment and evaluation of real job applicants and effectively rule out the

possibility that female applicants are indeed performing at a lower level than male candidates.

Interrupting, pausing, and speech duration are some but not all of the behaviors that matter in this setting. I selected this set of items because they were theory-relevant and it was relatively easy to measure them reliably and systematically. The behaviors I focused on, measured, and analyzed represent a small part of what presumably happens in interaction and impacts evaluation. For example, it would be interesting to measure judges' tone of voice. A qualitative observation is that male judges used a louder tone of voice when interviewing female applicants. In this setting a raised tone of voice directed at applicants would constitute another type of sanction similar to interruptions. It would have been interesting to see if judges' raised tone of voice correlated to applicants' pausing and other behaviors. This analysis was not done because noise conditions varied considerably across recordings.

The most important lesson to be gathered from this work is that existing inequalities are extremely hard to detect. The differential treatment of male and female applicants discussed here may be but a small portion of what really happens. With this I want to emphasize that, individually these differences in the treatment of men and women may not seem serious but, in the aggregate they have a very important cumulative effects. Receiving an underserved interruption may not change the course of the exam or the applicant's mindset. But if women feel that judges keep interrupting them, fidget, raise their tone of voice and so forth, this host of behaviors will definitely impact the female applicant's performance and reveals the judge's interpretation and evaluation of such performance.

Finally, the single most disturbing finding is that the objective quality of female applicants' performances does not impact their chances of passing the exam. One would think that the high quality answers would lead to higher scores. This is true for male applicants but not females. As DST argues, judges see quality as indicative of ability in male not female applicants. These findings align well with a recent study by Thomas-Hunt and Phillips (2004) where the authors found that women are often penalized when they possess the same expertise that men have (Thomas-Hunt & Phillips 2004).

This research is unique in that it uses very detailed data to document and examine an actual hiring process. This context is exceptional in that: (1) is accessible for direct observation and data collection. (2) the event of interest (i.e. exam) repeats sufficiently so as to evaluate theory-driven claims statistically, and (3) exams are fairly structured. which deems the lack of strict controls less problematic.

My findings suggest subtle but important evidence of judges' partiality in favor of male applicants. The subtlety of such corroborations indicates that a major effort needs to be made to identify sources of bias in real-world environments. Inequalities that surface in the course of live interactions are difficult to pin down. Whatever behavior one identifies may be attributed to employers or employees. In order adjudicate between supply and demand-type of explanations, unbiased indicators such as question difficulty and answer quality need to be gathered and analyzed jointly. As I argued in this project, the evidence of bias found illustrate the hardly detectable nature of these mechanisms. My findings align well with what Ridgeway identifies as the cause of the glass ceiling: "the

performance expectations and legitimacy reactions created by gender status beliefs create multiple, nearly invisible nets of comparative devaluation that catch women as they push forward to achieve positions of leadership and authority and slow them down compared to similar men" (Ridgeway 1994).

WORKS CITED

Anker, R. 1997. Theories of Occupational Segregation by Sex: An Overview. INTERNATIONAL LABOUR REVIEW 136:315.

Arrow, K. J. 1973. Social Responsibility and Economic Efficiency. PUBLIC POLICY 21:303-317.

Balkwell, J. W. 1991. From Expectations to Behavior: an Improved Postulate for Expectation States Theory. AMERICAN SOCIOLOGICAL REVIEW 56:355-369.

Becker, G. S. 1985. Human Capital, Effort, and the Sexual Division of Labor. JOURNAL OF LABOR ECONOMICS, 3:S33-S58

Berger, J., H. Fisek, R. Norman, and M. Zelditch, Jr. 1977. Status Characteristics and Social Interaction. New York: Elsevier.

Berger J., M. Zelditch, and B. Cohen. 1972. Status Characteristics and Social Interaction. AMERICAN SOCIOLOGICAL REVIEW 37:241.

Berger, J., S. Rosenholz, and M. Zeldtich. 1980. Status Organizing Processes. ANNUAL REVIEW OF SOCIOLOGY *6*, 479–508.

Biernat, M., and K. Fuegen. 2001. Shifting Standards and the Evaluation of Competence: Complexity in Gender-Based Judgment and Decision Making. JOURNAL OF SOCIAL ISSUES 57:707-724.

Biernat, M., and D. Kobrynowicz. 1997. Gender and Race Based Standards of Competence: Lower Minimum Standards but Higher Ability Standards for Devalued Groups. JOURNAL OF PERSONALITY AND SOCIAL PSYCHOLOGY 72:544-557.

Blair, I., and M.R. Banaji. 1996. Automatic and Controlled Processes in Stereotype Priming. JOURNAL OF PERSONALITY AND SOCIAL PSYCHOLOGY 70:1142-1163.

Blau, F. D., and M. Kahn. 2006. The Gender Pay Gap: Going, Going, … But Not Gone. In The Declining Significance of Gender?, Blau F., Brinton M., Grusky, D. (Eds.) New York: Russell Sage Foundation

Brewer, M., and R. Brown. 1998. Intergroup relations. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), Handbook of Social Psychology (pp. 554-594). New York: McGraw-Hill.

Brewer, M., and L. N. Lui. 1989. The Primacy of Age and Sex in the Structure of Person Categories. SOCIAL COGNITION 7:262-274.

Castilla, E. J. 2005. Social Networks and Employee Performance in a Call Center. AMERICAN JOURNAL OF SOCIOLOGY 110:1243-1283

Conway, M., M.T. Pizzamiglio, and L. Mount. 1996. Status, Communality, and Agency: Implications for Stereotypes of Gender and Other Groups. JOURNAL OF PERSONALITY AND SOCIAL PSYCHOLOGY 71:25-38.

Correll, S.J., and C. L. Ridgeway. 2003. Expectation States Theory, in Handbook of Social Psychology. New York: Kluwer Academic / Plenum Publishers.

Correll, S.J. 2001. Gender and the Career Choice Process: The Role of Biased Self-Assessments. AMERICAN JOURNAL OF SOCIOLOGY 106:1691-1730.

— 2004. Constraints into Preferences: Gender, Status, and Emerging Career Aspirations. AMERICAN SOCIOLOGICAL REVIEW 69:93-113.

Correll, S.J., and S. Benard. 2006. Biased Estimators? Comparing Status and Statistical Theories of Gender Discrimination, pp. 89-116 in Shane R. Thye and Edward J. Lawler (Eds). Social Psychology of the Workplace (Advances in Group Process Volume 23). New York: Elsevier.

Crespo-Montes, L. F. 2004. Los Administradores Civiles del Estado. Madrid: Instituto Nacional de Administración Pública.

Davison, H. K., and Burke, M. J. 2000. Sex Discrimination in Simulated Employment Contexts: A Meta-analytic Investigation. JOURNAL OF VOCATIONAL BEHAVIOR, 56:225-248.

Deaux, K., and T. Emswiller. 1974. Explanations of Successful Performance on Sex-Linked Tasks: What Is Skill for Male Is Luck for Female. JOURNAL OF PERSONALITY AND SOCIAL PSYCHOLOGY 29:80-85.

Deaux, K., and M. Kite. 1987. Thinking about Gender, in Analyzing Gender: A Handbook of Social Science Research, B. Hess, and M. Ferree (Eds). Newbury Park. CA: Sage.

Dodge, K.A., F. D. Gilroy, and L. M. Fenzel. 1995. Requisite Management Characteristics Revisited: Two Decades Later. JOURNAL OF SOCIAL BEHAVIOR AND PERSONALITY, 10(6), 253-264.

Dovidio, J. F., and Ellyson, S. L. 1985. Patterns of Visual Dominance Behavior in Humans, in S. Ellyson, and J. Dovidio (Eds.). Power, Dominance, and Nonverbal Behavior, pp. 129-150, New York: Springer-Verlag.

Eakins, B., and Eakins, R. G. 1978. Sex differences in human communication. Boston: Houghton Mifflin.

Erickson, K . G. 1998. The Impact of Cultural Status Beliefs on Individual Task Performance in Evaluative Settings: A New direction in Expectation States Research. PhD dissertation, Department of Sociology, Stanford University.

Farkas, G., P. England, K. Vicknair, B. Kilbourne. 1991. Subordinated Groups, Cognitive Skill, Occupational Access, and Wage Discrimination. Pres. Ann. Meet. Res. Com. 28, International Sociological Association, Columbus, Ohio.

Fernandez, R.M. and M. L. Sosa. 2005. Gendering the Job: Networks and Recruitment at a Call Center. AMERICAN JOURNAL OF SOCIOLOGY 111:859-904.

Fiske, S. T. 1992. Thinking Is for Doing. JOURNAL OF PERSONALITY AND SOCIAL PSYCHOLOGY 63:877-889.

Fiske, S. T., and S. E. Taylor. 1991. Social Cognition. New York, NY: McGraw Hill Book Company.

Foschi, M. 1989. Status Characteristics. Standards and Attributions, in Sociological Theories in Progress, J. Berger, M. Zelditch, and B. Anderson (Eds), pp. 58-72, Newbury Park. CA: Sage.

Foschi, M. 1996. Double Standards in the Evaluation of Men and Women. SOCIAL PSYCHOLOGY QUARTERLY 59:237-254.

— 2000. Double Standards for Competence: Theory and Research. ANNUAL REVIEW OF SOCIOLOGY 26:21-42.

Foschi, M., L. Lai., and K. Sigerson. 1994. Gender and Double Standards in the Assessment of Job Applicants. SOCIAL PSYCHOLOGY QUARTERLY 57:326-339.

Goldin, C., and C. Rouse. 2000. Orchestrating Impartiality: the Impact of 'Blind' Auditions on Female Musicians. AMERICAN ECONOMIC REVIEW 90:715-741.

Heilman, M. E. 1983. Sex Bias in Work Settings: The Lack of Fit Model, in B. Staw, and L. Cummings (Eds.) Research in organizational behavior 5:269-298. Greenwich CT: JAI.

Heilman, M. E. 2001. Description and Prescription: How Gender Stereotypes Prevent Women's Ascent Up the Organizational Ladder. JOURNAL OF SOCIAL ISSUES 57:657-574.

Heilman, M. E., C. J. Block, and R. F. Martell, R. F. 1995. Sex Stereotypes: Do they Influence Perceptions of Managers? JOURNAL OF SOCIAL BEHAVIOR AND PERSONALITY 10, 237–252.

Heilman, M. E., and E. J. Parks-Stamm. 2007. Gender Stereotypes in the Workplace: Obstacles to Women's Career Progress. SOCIAL PSYCHOLOGY OF GENDER 24: 47-77.

Klimoski, R. and L. Inks. 1990. Accountability Forces in Performance Appraisal. ORGANIZATIONAL BEHAVIOR AND HUMAN DECISION PROCESSES 45:194-208.

Kollock, P., P. Blumstein, and P. Schwartz. 1985. Sex and Power in Interaction - Conversational Privileges and Duties. AMERICAN SOCIOLOGICAL REVIEW 50:34-46.

Kunda, Z., L. Sinclair, and D. Griffin. 1997. Equal Ratings but Separate Meanings: Stereotypes and the Construal of Traits. JOURNAL OF PERSONALITY AND SOCIAL PSYCHOLOGY 72:720-734.

Lafrance, M. 1992. Gender and Interruptions - Individual Infraction or Violation of the Social-Order. PSYCHOLOGY OF WOMEN QUARTERLY 16:497-512.

Landsberger, H. A. 1958. Hawthorne Revisited. Ithaca. NY: Cornell University.

Lockheed, M. E., and K. P. Hall. 1976. Conceptualizing Sex as a Status Characteristic: Application to Leadership Training Strategies. JOURNAL OF SOCIAL ISSUES *32*, 111-124.

Lovaglia, M. J., J. W. Lucas, J. A. Houser, S. R. Thye, and B. Markovsky. 1998. Status Processes and Mental Ability Test Scores. American Journal of Sociology 104:195-228.

Lyness, K. S. 2002. Finding the Key to Executive Suite: Challenges for Women and People of Color, in R. Silzer (Ed.) The 21st Century Executive: Innovative Practices for Building Leadership at the Top 229-273. San Francisco, CA: Jossey-Boss.

Lyness, K. S., and M. E. Heilman. 2006. When Fit is Fundamental: Performance Evaluation and Promotions of Upper-level Female and Male Managers. JOURNAL OF APPLIED PSYCHOLOGY 91:777-785

Marche, T.A., and C. Peterson. 1993. The Development and Sex-Related Use of Interruption Behavior. HUMAN COMMUNICATION RESEARCH 19:388-408.

Meeker, B. F., and P. A. Weitzel-O'Neil. 1977. Sex Roles and Interpersonal Behavior in Task-oriented Groups. AMERICAN SOCIOLOGICAL REVIEW 42, 91-105.

Mueller, C.W., M. Mulinge, and J. Glass. 2002. Interactional Processes and Gender Workplace Inequalities. SOCIAL PSYCHOLOGY QUATERLY 65:163-185.

Nieva, V. G., and B. A. Gutek. 1980. Sex Effects on Evaluation. ACADEMY OF MANAGEMENT REVIEW, 5:267-276

Neumark, D. M. 1996. Sex Discrimination in Restaurant Hiring: An audit Study. QUATERLY JOURNAL OF ECONOMICS 111:915-941

O'Leary, V. E., and B. S. Wallston. 1982. Women, Gender, and Social Psychology. REVIEW OF PERSONALITY AND SOCIAL PSYCHOLOGY 2:9-43.

O'Leary, V. E. 1974. Some Attitudinal Barriers to Occupational Aspirations in Women. PSYCHOLOGICAL BULLETIN 81:809-826.

Perry, E. L., A. Davis-Blake, and C. T. Kulik. 1994. Explaining Gender-Based Selection Decisions: a Synthesis of Contextual and Cognitive Approaches. ACADEMY OF MANAGEMENT REVIEW 19:786-820.

Petersen, T., and L. A. Morgan. 1995. Separate and Unequal - Occupation Establishment Sex Segregation and the Gender Wage Gap. AMERICAN JOURNAL OF SOCIOLOGY 101:329-365.

Phelps, E. S. 1972. Statistical Theory of Racism and Sexism. AMERICAN ECONOMIC REVIEW 62:659-661.

Pugh, M. D., and R. Wahrman. 1983. Neutralizing Sexism in Mixed-Sex Groups - Do Women Have to Be Better Than Men. AMERICAN JOURNAL OF SOCIOLOGY 88:746-762.

Reskin, B. 1993. Sex Segregation in the Workplace. ANNUAL REVIEW OF SOCIOLOGY 19:241-270.

Ridgeway, C. L. 1987. Nonverbal Behavior, Dominance, and the Basis of Status in Task Groups. AMERICAN SOCIOLOGICAL REVIEW 52.

Ridgeway, C. L. 1997. Interaction and the Conservation of Gender Inequality: Considering Employment. AMERICAN SOCIOLOGICAL REVIEW 62:218-235.

— 2001. Gender, Status, and Leadership. JOURNAL OF SOCIAL ISSUES 57:637-655.

Ridgeway, C. L., and J. Berger. 1986. Expectations, Legitimation, and Dominance Behavior in Groups. AMERICAN SOCIOLOGICAL REVIEW 51, 603-617.

Ridgeway, C. L., J. Berger, and L. Smith-Lovin. 1985. Nonverbal Cues and Status - an Expectation States Approach. AMERICAN JOURNAL OF SOCIOLOGY 90:955-978.

Ridgeway, C. L., and S. J. Correll. 2000. Limiting Inequality through Interaction: The End(S) of Gender. CONTEMPORARY SOCIOLOGY A JOURNAL OF REVIEWS 29:110-120.

— 2004. Unpacking the Gender System - a Theoretical Perspective on Gender Beliefs and Social Relations. GENDER & SOCIETY 18:510-531.

Ridgeway, C. L., and K. G. Erickson. 2000. Creating and Spreading Status Beliefs. AMERICAN JOURNAL OF SOCIOLOGY 106:579-615.

Ridgeway, C. L., C. Johnson, and D, Dikema. 1994. External Status. Legitimacy, and Compliance in Male and Female Groups. SOCIAL FORCES 72:1051-1077.

Ridgeway, C. L., and L. Smith-Lovin. 1999. The Gender System and Interaction. ANNUAL REVIEW OF SOCIOLOGY 25:191-216.

Rudman, L. A., and P. Glick. 2001. Prescriptive Gender Stereotypes and Backlash Toward Agentic Women. JOURNAL OF SOCIAL ISSUES 57:743-762.

Simonson, I., and P. Nye. 1992. The Effect of Accountability on Susceptibility to Decision Errors. ORGANIZATIONAL BEHAVIOR AND HUMAN DECISION PROCESSES 51:416-446.

Smith-Lovin, L., and C. Brody. 1989. Interruptions in Group Discussions - the Effects of Gender and Group Composition. AMERICAN SOCIOLOGICAL REVIEW 54:424-435.

Stangor, C., L. Lynch, C. M. Duan, and B. Glass. 1992. Categorization of Individuals on the Basis of Multiple Social Features. JOURNAL OF PERSONALITY AND SOCIAL PSYCHOLOGY 62:207-218.

Steele, C. M. 1997. A Threat in the Air - How Stereotypes Shape Intellectual Identity and Performance. AMERICAN PSYCHOLOGIST 52:613-629.

Steinpreis, R. E., K. A. Anders, and D. Ritzke. 1999. The Impact of Gender on the Review of the Curricula Vitae of Job Applicants and

Tenure Candidates: A National Empirical Study. SEX ROLES 41:509-528.

Stewart, P. A., and J. C. Moore. 1992. Wage Disparities and Performance Expectations. SOCIAL PSYCHOLOGY QUARTERLY 55.78-85.

Thomas-Hunt, M. C., and K. W. Phillips. 2004. When What You Know Is Not Enough: Expertise and Gender Dynamics in Task Groups. PERSONALITY AND SOCIAL PSYCHOLOGY BULLETIN 30:1585-1598.

West, C., and D. Zimmerman. 1983. Small Insults: A Study of Interruptions in Cross-Sex Conversations between Unacquainted Persons, in B. Thorne, C. Kramarae, and N. Henley (Eds.) Language, Gender, and Society. Rowley, M.A.: Newbury House.

Williams, J. E., and D. L. Best. 1990. Measuring Sex Stereotypes: A Multinational Study. Beverly Hills: Sage.